Large Scale Data Analytics of User Behavior for Improving Content Delivery

Athula Balachandran

CMU-CS-14-142

December 2014

School of Computer Science
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Srinivasan Seshan, Co-Chair
Vyas Sekar, Co-Chair
Hui Zhang
Peter Steenkiste
Aditya Akella, University of Wisconsin-Madison

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research was sponsored by the National Science Foundation under grant numbers CNS-0721857, CNS-0905277, and CNS-1040801; and the U.S. Army Research Office under grant number W911NF0910273.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

Keywords: data analytics, machine learning, user behavior, user experience, content delivery, peer-to-peer, video streaming, web browsing

To Achan and Amma

Abstract

The Internet is fast becoming the de facto content delivery network of the world, supplanting TV and physical media as the primary method of distributing larger files to ever-increasing numbers of users at the fastest possible speeds. Recent trends have, however, posed challenges to various players in the Internet content delivery ecosystem. These trends include exponentially *increasing traffic volume*, *increasing user expectation* for quality of content delivery, and the ubiquity and *rise of mobile traffic*.

For example, exponentially increasing traffic—primarily caused by the popularity of Internet video—is stressing the existing Content Delivery Network (CDN) infrastructures. Similarly, content providers want to improve user experience to match the increasing user expectation in order to retain users and sustain their advertisement-based and subscription-based revenue models. Finally, although mobile traffic is increasing, cellular networks are not as well designed as their wireline counterparts, causing poorer quality of experience for mobile users. These challenges are faced by content providers, CDNs and network operators everywhere and they seek to design and manage their networks better to improve content delivery and provide better quality of experience.

This thesis identifies a new opportunity to tackle these challenges with the help of big data analytics. We show that large-scale analytics on user behavior data can be used to inform the design of different aspects of the content delivery systems. Specifically, we show that insights from large-scale analytics can lead to better resource provisioning to augment the existing CDN infrastructure and tackle increasing traffic. Further, we build predictive models using machine learning techniques to understand users' expectations for quality. These models can be used to improve users' quality of experience. Similarly, we show that even mobile network operators who do not have access to client-side or server-side logs on user access patterns can use large-scale data analytics techniques to extract user behavior from network traces and build machine learning models that help configure the network better for improved content delivery.

Acknowledgments

This dissertaton marks the end of a very long chapter in my life, and serves as the stepping stone to another exciting and maybe a tad intimidating chapter. I've been in "school" of some form for as long as I can remember, going on 24 years now, and this disseration is my cue to say goodbye to school and to step into the "real world". While working on this research, I have learned so much from my mentors, colleagues, and friends, and I feel strangely confident that this journey has prepared me for any challenge I may face in work or life.

I'm lucky to have such a stellar group of researchers on my thesis committee. My advisor, Srini Seshan, was the best advisor I could have hoped for to guide me through this PhD. As a green first-year, Srini showed me the ropes on the hot topics in networking, and gave me the freedom to work on what I found interesting. Srini sees the field from a vantage point that few others do, and he was able to teach me which trees yield fruits and which ones do not. His positive attitude gave me the confidence to keep going after paper rejections.

Vyas Sekar has been involved in my research since the day I arrived at CMU—first as a mentor, then a collaborator, and now as my co-advisor. Vyas was always available to help with writing or ideas. With his razor-sharp intuition, Vyas would have a solution to any problem I faced, and he'd be ready to jump on Skype or Hangouts to help me day or night. I also have Vyas to thank for helping me improve my shoddy writing over the years.

I'm fortunate that I've been able to pick Aditya Akella's brain during our several calls; his suggestions largely paved the way for our IMC 2013 paper. I'm also immensely grateful to Hui Zhang; he trusted a novice PhD student with production data at Conviva, which was the keystone of the IMC and SIGCOMM papers, and indeed the turning point for my PhD. Peter Steenkiste kept tabs on my research, encouraged me to present my work during Tuesday seminars, and more lately, helped greatly improve this dissertation with his detailed feedback.

Outside of my thesis committee, I'm fortunate to have met and collaborated with several luminaries in the field. In my first summer at Intel Research, Nina Taft, Kevin Fall, and Gianluca Iannacone took me under their wing. Ion Stoica gave guidance for the QoE work during my visit to Conviva. I am also indebted to folks at Conviva especially Dilip, Xi, Dima and Jibin for answering several questions about the Conviva data and cluster. Jeff Pang, my mentor during my internship at AT&T Labs, guided me through the mobile QoE work that became our Mobicom paper. I am also indebted to my other co-authors: Ashok Anand, Emir Halevopic, Shobha Venkataraman, He Yan and Vaneet Agarwal.

Gloomy winters and research troubles were more than neutralized by an amazing set of friends within and outside the department. My officemates Soonho, Dongsu, George, and Carol helped take my mind off work through numerous fun conversations. I'll also fondly remember chats with other GHC 7th floor denizens: Anvesh, Ankit, Mehdi, Sarah, Dana, Yair, Gabriel, Joao, David, Matt, Junchen, and Richard. I am also indebted to Deborah Cavlovich, Angela Miller and the rest of the excellent

staff and faculty at CMU for their support and help over the years. My housemates and neighbors made my life at home fun. Sunayana, Bhavana, Meghana, Anjali, Kriti and Ruta have been great company and have let me partake in their delicious homecooked food more times than I can remember. My friends from undergrad at CMU—Leela, Vivek, and Srivatsan—filled me up on insti news and memories. Though separated by longer distances, my undergrad buddies Nitya, Pranava and Deepa helped me let off steam over phone, Skype, or occasional visits.

Through my entire academic journey, my parents have encouraged me even when I did not believe in myself; this work is as much their effort as it is mine. My grandmother and uncle have been pillars of support from when I was young. My relatives and cousins in the US were my home away from home over the past five years. Over the past year, I'm also fortunate to have had the encouragement of my parents-in-law. I am extremely blessed to have an amazing companion. Having done a PhD in networking himself, Anirudh was very understanding of the ways of a PhD student. He was the source of my sanity, a sounding board for my random ideas and an excellent proof-reader for my drafts and slides.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Internet today is largely a content driven network. Starting from simple data transfer between two computers directly connected by a wire, the complexity of content delivery over the Internet has come a long way to include several complex applications such as adaptive video streaming, peer-to-peer file sharing, massively multiplayer online gaming, cloud storage, and cloud-based computation. Over the years, there have been several innovations to support the growth of content delivery, both in protocols used for delivering content, as well as in the infrastructure to support and improve new content delivery applications.

Notable protocol innovations include optical technology for very high speed connectivity at the physical level [26], high-speed variants of IEEE 802.3 Ethernet specification at the data-link layer [1], IP multicast at the network layer, specialized transport layer mechanisms such as SCTP [34] and DCCP [13] specifically for streaming media and video, and application layer protocols such as Real Time Streaming Protocols [32] and Dynamic Adaptive Streaming over HTTP [12]. There have also been several infrastructure innovations, especially in the design of content delivery systems. Beginning with caches placed in front of content servers to return frequently-requested content, content providers began distributing these caches globally close to the request origin using Content Distribution Networks (CDNs). Other content delivery system design innovations include peer to peer content delivery such as BitTorrent and its variants, and hybrid P2P-CDNs, where, in addition to the content servers, clients of the CDN also contribute content they have downloaded to peer clients [64].

These protocol and infrastructure innovations have greatly improved content delivery over the Internet making it very robust. Today, as a result, tens of millions of people consider the Internet to be a necessity. They work, bank, communicate, plan travel, find food, and seek entertainment using Internet-based services. However, with the ubiquity of Internet-connected devices, and with online services demanding high bandwidths and low latency, there are challenges faced by all players in the ecosystem. Users expect instant, high-quality connectivity to their services from any device, and the players in the content delivery ecosystem know that better performance is tightly correlated with higher revenues. But challenges such as increasing load on the network and heterogeneity of technologies especially in cellular data networks (UMTS, LTE) may not provide the best quality of experience to users at all times and on all devices. The players in the ecosystem know that it could be costly to not deal with these challenges ahead of time. For instance, a single second of downtime for a content delivery service like Google or Amazon costs

these services several hundred thousands of dollars in revenue [15, 72]

Today the main challenges faced by the content delivery ecosystem include:

1. *Exponentially increasing traffic:* Traffic over the Internet has been exponentially increasing over the past few years with predictions that it will quadruple by 2016 [9]. In 2011, 51% of the traffic on the Internet consisted of video. Market predictions suggest that more than 90% of the traffic on the Internet will be video in 2015. This development is hugely due to the very low costs of obtaining video over the Internet. In fact, we have come to a point where Internet video would replace traditional TV viewership. However, recent studies have shown signs of the CDN infrastructure being stressed by the increasing traffic [95]. This is placing an onus on the CDNs to distribute content efficiently. Some of the techniques to augment the existing infrastructure to handle the increasing load, that have received significant industry attention, include hybrid P2P-CDN design which combines the P2P and client-server models for content delivery [63, 64], and federated CDN models [48, 105].
2. *Increasing user expectation:* Another side effect of the decreasing costs of obtaining content, specifically video, over the Internet includes the rise of several content providers competing for users' attention. With multiple competitors in the market, user expectation for the quality of the content has been steadily growing [3]. Content providers also want to maximize user engagement in order for better gains from their advertisement-based or subscription-based business models. This has led to improving user experience as one of the primary goals of content delivery today. This trend has spawned several third-party optimization and analytics services that operate for optimizing user experience given the limited resources using techniques such as cross-CDN optimization.
3. *Rise of mobile:* Another major development in the past few years in the content delivery scenario is the rise of mobile traffic. Today, mobile traffic constitutes 28% of the overall Internet traffic with one in four visits to websites arising from a smartphone. Mobile web usage is estimated to increase eleven-fold between 2013 and 2016 with around 50 billion connected mobile devices on the Internet by 2020 [9]. However, today, cellular networks are slower than wireline networks primarily because the mobile architecture was not designed for the web. Recent studies on top websites showed that loading them via wireline connectivity lead to an average of 2 seconds when compared to 9 seconds via mobile connectivity [3]. Increasing user expectation is posing challenges to the cellular network operators to configure their networks by adding the right infrastructure in order to provide wireline compatible user experience over cellular networks.
4. *Complex Multiple Party Involvement:* The content delivery ecosystem today consists of several parties including content providers, CDNs, network operators, third party analytics and CDN optimization services etc. The increasing complexity and multiple party involvement makes it much harder to pinpoint problems in delivering content to users [85]. Harder still is to estimate whether a proposed feature or fix to a system or protocol will have a measurable impact on users or revenue. If a user complains about a video not loading on her smartphone, what can the content provider do? Perhaps the problem is the buffering time or the bitrate selected by the CDN or third-party optimization services. Or it could be caused by overload at the CDN server. It could otherwise be simply because the user's

device switched from a 4G to a 3G connection. Even if one or more of these issues are addressed, content providers, network operators, CDNs etc. traditionally have no way to know if the problem is widespread or if the fix had the desired impact.

With the increasing user base, increasing user expectation, and the diversity of access methods, content providers, CDNs, and network operators are faced with a very complex content delivery system. This thesis identifies a new opportunity to help tackle these challenges by using large scale data analysis on the data collected by the content providers, the CDN, and the network operator. Over the last decade, there have been significant advances in computer science particularly towards the development of big data analytics algorithms and systems. With the ability to collect and store fine-grained information about all aspects of this traffic, this opens up the unprecedented opportunity to explore the usage of big data analytics to not only help improve network protocols and designs of content delivery to meet the challenges, but also potentially incorporate these techniques and models to the run time operations of the content delivery systems.

The goal of this thesis is to identify interesting user behavior using large scale data analytics and transform it into actionable insights and models that can be used to improve various aspects of content delivery. Specifically we show that insights from large-scale analytics can lead to better resource provisioning to augment the existing CDN infrastructure and tackle increasing traffic. Further, we build predictive models using machine learning techniques to understand users' expectations for quality. These models can be used to improve users' quality of experience of users. Similarly, we show that even mobile network operators who do not have access to client-side or server-side logs on user access patterns can use large-scale data analytics techniques to extract user behavior from network traces and build machine learning models that help configure the network better for better content delivery.

1.1 Background and Scope

This section outlines the scope of our dissertation research in more detail. We first present the state of the art of the content delivery ecosystem in Section 1.1.1, followed by the opportunity presented by big data analytics in Section 1.1.2. We then present the scope of the thesis in Section 1.1.3.

1.1.1 Content Delivery Ecosystem

We begin with a brief overview of the different players in the content delivery ecosystem in the Internet today. Each of these players have access to rich data that can be used towards improving content delivery.

- **Content providers** encompass a wide variety of media and e-commerce players who provide content on the Internet primarily for revenue. These include news websites (e.g., CNN), social networking websites (e.g., Facebook, Yelp), and also video providers (e.g., HBO, ABC). Content providers want to maximize their revenues from subscription-based and advertisement-based business models while trying to minimize content distribution costs. To this end, content providers have business arrangements with CDNs (e.g., Akamai, Limelight) to distribute their content across different geographical locations. Similarly,

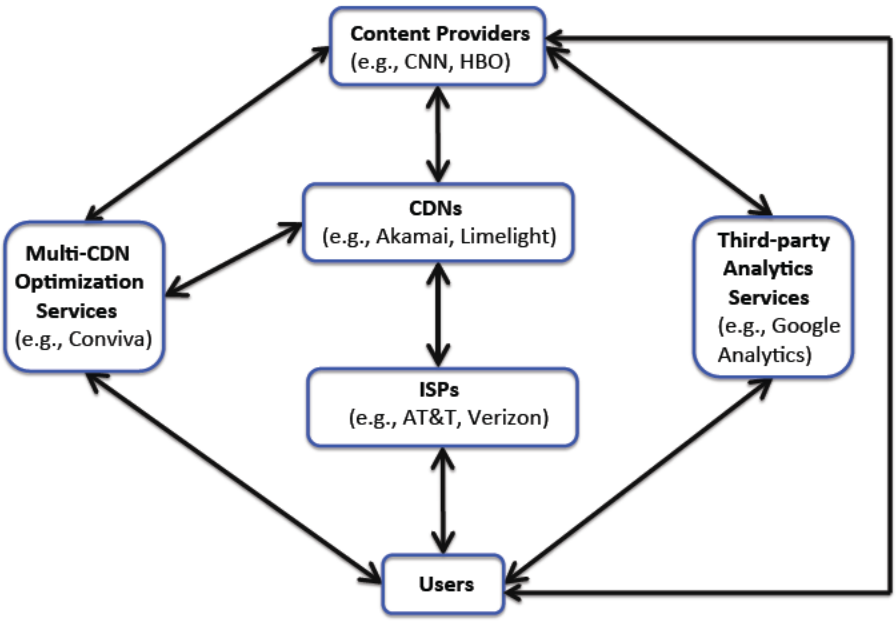


Figure 1.1: Overview of the Internet content delivery ecosystem

more recently they also have contracts with third-party analytics services (e.g., Google Analytics, Ooyala [28]) and optimization services (e.g., Conviva [11]) to understand and improve user experience.

- **Content Distribution Networks** (e.g., Akamai, Limelight) consist of distributed system of servers allocated across different geographical regions for serving content to end users with high performance and availability. CDNs provide content providers a cost-effective mechanism to offload content from their infrastructure. Hence CDNs need to allocate their resources (e.g., server and bandwidth capacity) efficiently across user population. CDNs aim to design their delivery infrastructure to minimize their delivery costs while maximizing their performance. Towards this end there have been many studies and proposals on efficient design of the CDN infrastructure. Although CDNs primarily serve content using dedicated servers operated by them, more recently there have been proposals for other designs including hybrid models that make use of peer-to-peer mechanisms on user-owned devices and also federation across multiple CDNs. CDNs collect a large amount of logs daily on user behavior. Tailoring CDN design based on the user behavior to improve content delivery with minimal costs is an interesting problem faced by CDNs.
- **Internet Service Providers (ISPs)** form the backbone of the Internet by delivering the content from the CDNs and content providers to the end users. Traffic on the Internet has been increasing exponentially over the years. In particular, with the advent of smartphones and new wireless technologies such as 3G and 4G, mobile traffic is on the rise. But, unlike the other players, ISPs do not have access to detailed client-side or server-side logs making it more challenging to extract user behavior information from network traces alone.

However, extracting user behavior information from network traces can help ISPs can use this data towards improving content delivery by configuring their network better.

- **Cross-CDN optimization services** help content providers work with multiple CDNs towards delivering content for better resilience. Recent studies have also argued that cross-CDN optimization can lead to improved content delivery [95]. There are also commercial players in the market (e.g., Conviva) that offer cross-CDN optimization services for content providers, especially in Internet video providers. These services also have access to user behavior at a fine grained level at a large-scale and make real-time decisions on which CDNs to serve a content based on current network conditions. Such optimizations are towards improving user experience while mainting low content-delivery costs. Similarly, there are also several **third party analytics services** (Google Analytics, Ooyala) that collect user access logs information at a large-scale to translate them into insights towards improving revenue for the user at low content delivery costs.
- **Users** are ultimately the source of all revenue and the sink for all content produced by this ecosystem. Users prefer services that give them a better cost-experience tradeoff and hence content providers need to deliver the best possible quality of experience to the users at the minimum cost. At the same time, we now have the ability to collect, store and analyze fine-grained access patterns and behavior from the users. This information, even at an aggregate level can help the content delivery system to minimize costs by provisioning resources appropriately and also improve individual user's quality of experience by potentially even personalizing content delivery based on their preferences.

With the increasing traffic on the Internet, each player in the ecosystem now can capture a lot of data on user behavior and preferences. In this thesis we argue that this data can be used towards improving content delivery.

1.1.2 Big Data Analytics

Big data analytics is now extensively used in fields of computer science such as recommendation systems, search and information retrieval, computer vision and image processing, and is making its foray into the real world in terms of business intelligence, healthcare and supply chain analysis. It is also used even within the domain of networks in areas such as network security.

Several technology innovations in the past decade were essential in being able to analyze massive volumes of data. The MapReduce framework [22] is perhaps the innovation that heralded the area of big data analytics, and open-source versions of MapReduce such as Hadoop [18] and the distributed HDFS [19] filesystem allow researchers to rapidly gather insights from more data that can fit on any single machine. Hadoop, Hive [20] and recent advancements such as Spark [36] make short work of analyzing massive quantities of data. Keeping up with the infrastructure developments, there have been algorithms and libraries that are specifically suited to data mining and machine learning, ranging from the more traditional tools such as Weka [42] and Scikit-learn [33] to tools built for big data such as Graphlab [17].

In our work, we make heavy use of the Hadoop and HDFS infrastructure to process logs from various sources. Hive helped us easily interact with the data.

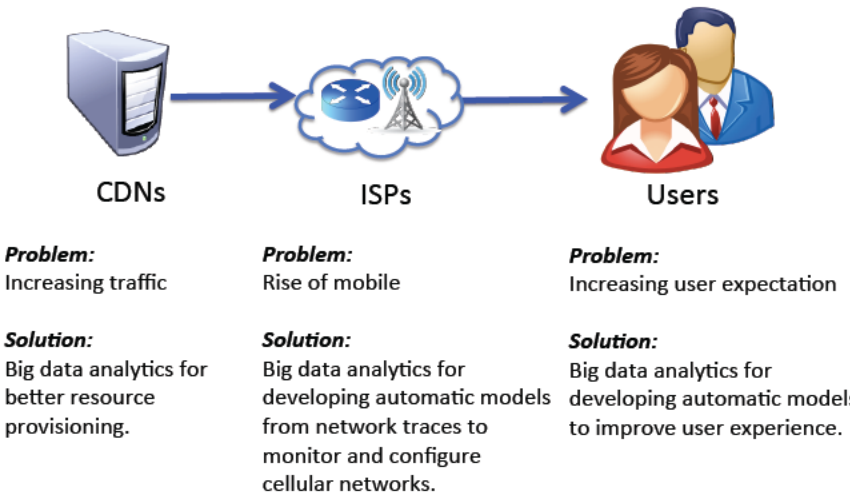


Figure 1.2: Flow of information during content delivery from CDNs to ISPs to Users. We look at how we can use large-scale data analytics to help improve content delivery at each point in the flow.

1.1.3 Thesis Scope

There are several ways in which big data analytics could potentially help improve content delivery. In this dissertation the main focus is on how big data analytics on user behavior can be of use to improve content delivery over the Internet. Big data analytics can also be used to predict network throughput, to build better control plane and data plane mechanisms and hence improve content delivery. These are interesting topics for future work and are discussed in more detail in Chapter 5.

1.2 Thesis Statement and Approach

Big data analytics tools can be used to improve content delivery not just by informing system design decisions, but also by building automatic models that can be directly used in decision processes. Based on this idea, this thesis argues the following:

> *It is possible for different players to use **big data analytics** on user behavior data for learning **predictive models** that can be used to **improve content delivery** even when the data collected does **not explicitly** contain user behavior information.*

To support this claim, this thesis shows initial promise on how big data analytics can be used to help tackle the major challenges faced by content delivery today—increasing traffic, increasing user expectation, and rising mobile traffic. Figure 1.2 gives a high level overview of the thesis approach. We performed large-scale studies conducted on data from three different perspectives:

Study	Big data analytics	Improve content delivery	Predictive models	Not explicit Data
CDN resource management	✓	✓	✗	✗
Internet video QoE	✓	✓	✓	✗
Cellular web QoE	✓	✓	✓	✓

Table 1.1: Summary of contributions of each study in the thesis

1. *Inform CDN resource management:* We first look at how we can use large scale data analytics on user behavior data for gaining insights that can be used to inform system design changes for handling the ever-increasing traffic at CDNs. With exponentially increasing traffic, CDNs wish to minimize bandwidth costs while improving performance in terms of availability. This implies that they must provision and select servers appropriately. Most commonly, CDNs serve content from dedicated servers. However, there have been recent proposals to improve this scheme using hybrid P2P CDNs and federated CDNs. We show that it possible to analyze data already collected from existing CDNs to evaluate the advantages of these new proposals. We also observe several interesting user behavior patterns that have important implications to these designs. Using aggregate client access logs, we analyze proposals to improve provisioning that have seen significant industry traction in the recent times. This study shows that even simple data analytics can be very useful for improving content delivery.
2. *Develop Internet Video Quality of Experience (QoE) models:* Next we look at the problem of increasing user expectation for quality of content. Content providers are very keen on understanding and improving users' QoE since it directly affects their revenue. For instance, previous studies in Internet video viewing have shown that one percent increase in buffering can lower viewing time by as much as three minutes [72]. We use big data analytics, particularly machine learning algorithms to build predictive models that capture users QoE. Using simulation studies, we also show that using this model to pick the best bitrate and CDN server for a video session can lead to as much as 20% improvement in user engagement. This study shows that in addition to informing system design decisions, big data analytics can also be used to build predictive models that can potentially be directly used in systems.
3. *Model Cellular Web Browsing Quality of Experience:* Although QoE for various content delivery applications over the Internet has been improving over the years, QoE in certain spaces are still not as good. Cellular networks do not provide good QoE even for simple applications such as web browsing. Further, unlike other players in the ecosystem cellular network operators do not have access to detailed server-side and client-side logs of user behavior. We show that machine learning algorithms can be used to extract user behavior information from network traces and build predictive models for web browsing QoE from this extracted user behavior information. We show that these models can be used by network operators to monitor and configure their networks for better QoE. This study shows that big data analytics can be useful even in scenarios where the player does not have explicit user behavior information.

Table 1.1 summarizes the above three studies and their main contributions.

1.3 Thesis Contributions

To summarize, we make three main contributions in this thesis:

1. This dissertation shows that even simple data analytics can be very useful for informing system design decisions towards improving content delivery.
2. We show that in addition to informing system design decisions, data analytics particularly machine learning algorithms can be used to build predictive models that can potentially be used to improve content delivery.
3. In scenarios where players do not have access to detailed server-side or client-side logs (e.g., ISPs), we show that big data analytics can be used to extract user behavior information from inexplicit data such as network traces. Such information can be very useful for players to understand and improve content delivery.

1.4 Dissertation Outline

The remainder of the dissertation is organized as follows:

1. In Chapter 2, we present a study based on large-scale data analytics to inform provisioning decisions at the CDNs. Recent studies have shown that increasing traffic is stressing the current CDN infrastructure. More than 90% of this traffic consists of video sessions. Using large-scale data analytics on over 30 million video sessions, we identify several user access patterns that have implications on CDN provisioning. We also analyze the impact of our findings on two emerging strategies to augment the existing infrastructure (hybrid P2P-CDNs and federated CDNs) that have received significant attention from the industry recently.
2. In Chapter 3, we present an approach to use large scale analytics and machine learning to build models that capture users' quality of experience for Internet video. Improving users' quality of experience is crucial to sustain the advertisement and subscription based revenue models that drive different content delivery services. We build a machine learning model to capture the relationship between different quality metrics that content providers have control over to improve users' engagement. We also show that a delivery infrastructure that uses this model can potentially achieve better user QoE.
3. In Chapter 4, we present an approach to use machine learning to build models that characterize the impact of network parameters on user engagement in cellular networks. Cellular networks are not as well designed as their wireline counterparts. However, users expect high quality of experience. Also, cellular network operators, unlike other players, do not have access to detailed client-side and server-side logs on user behavior. We design and test machine learning approaches that can be used to extract user behavior metrics from network traces and also build models that capture the relationship between radio network

metrics to user experience. These models can be used by network operators to help configure their network to prioritize the improvement of factors that have the most impact on user experience.

4. In Chapter 5, we summarize the contributions of this thesis, list the lessons learnt and outline some of the open problems for future work.

Chapter 2

Large-Scale Data Analytics for CDN Resource Management

Traffic on the Internet has been steadily growing, and it is predicted to quadruple by 2018 [9]. Dealing with exponentially increasing traffic volumes is a significant challenge for different players in the content delivery ecosystem. Video accounts for a large fraction of the traffic on the Internet today, and its share is growing over time. In 2011, around 51% of Internet traffic was video [9], and market predictions suggest that video will account for over 90% of the traffic on the Internet in 2015. There are already signs that the CDN infrastructure is being stressed [94, 95] by the increasing traffic and this has placed the onus on CDNs for managing their resources more efficiently. In this chapter, we show that large scale data analytics can be used to inform CDN resource management challenges caused by exponentially increasing traffic volumes. Since video forms the largest fraction of Internet traffic, our study in this chapter focuses on Internet video viewing workload.

Hybrid P2P-CDNs and telco-CDN federation are two CDN infrastructure augmentation strategies to alleviate the stress caused by increasing traffic. These two approaches have received significant industry attention recently.

- Telco-CDN federation is based on the recent development amongst various CDNs operated by telecommunication companies to federate by interconnecting their networks and compete directly with the traditional CDNs [8, 48, 70, 105]. This would enable users to reach CDN caches that are closer. Interconnecting resources across telco-CDNs would also ensure better availability and will benefit the participating ISPs in terms of provisioning costs [8].
- A hybrid strategy of serving content from dedicated CDN servers using P2P technology (e.g., [63, 64]) has been around for a while in the research literature but has only recently seen traction in the industry [4, 118]. A hybrid P2P-CDN approach would provide the scalability advantage of P2P along with the reliability and manageability of CDNs.

Given that several industry efforts and working groups are underway for both these approaches [4, 48, 70, 105, 118], it is crucial to analyze the potential benefits that these CDN augmentation strategies can offer for Internet video workloads. Our main contribution in this chapter is in using large-scale data analytics for identifying video access patterns that have sig-

nificant implications to these two strategies and analyzing the potential benefits of these two strategies. To the best of our knowledge, there has not been any previous large-scale study on the benefits of federated telco-CDN infrastructures. While there is prior work on analyzing the benefits of P2P augmentation, these were done long before Internet video became mainstream [63, 64], and hence were ahead of their times. Moreoever, the significant improvement in big data analytics approaches and the ability to collect large amounts of data puts us in a better position to do this study today. We leverage on these to revisit the benefits of P2P augmentation on today's traffic and suggest new improvements.

Using a dataset of around 30 million VOD and live sessions collected over two months from viewers across the United States, we identify several video viewing patterns that have implications to these two designs including:

- **Regional interest:** Typically, we observe significant population induced difference in load across different regions (e.g., US East coast, US West coast, Mid-West). But, for live events with regional biases like a local team playing a match, we observe significantly skewed access rates from regions that exhibit low load in the typical case.
- **Temporal shift in peak load:** We observe strong diurnal effects in access patterns and also confirm temporal shifts between regions in the demand for VOD objects using cross-correlation analysis. The temporal shift in access pattern is caused by time zone differences. The video access load peaks at around 8pm local time for each region.
- **Evolution of interest:** We observe that peak demand for VOD objects occur on the day of release and the decay in demand in the subsequent days can be modeled using an exponential decay process. Interestingly, overall user viewing patterns are very different across genres (e.g., TV series, reality shows, news shows). For example, decay rates of news shows are much higher than TV series episodes. Also, TV series episodes have highly predictable and stable demand from week to week (i.e., across successive episodes).
- **Synchronized viewing patterns:** While we expect synchronous viewing behavior for live video, we unexpectedly observe synchrony in the viewership of VOD objects. This is especially true for popular shows during the peak demand period (e.g., evening of the day of release of the show).
- **Partial interest in content:** We reconfirm prior observations that users watch only part of the video during a session [50, 76]. For instance, in the case of VOD, a significant fraction of the viewers typically watch only the first 10 minutes of the video before quitting. We observe that around 4.5% of the users are "serial" early-quitters (analogous to channel surfing) while 16.6% of the users consistently watch videos to completion.

We develop simple models to capture the deployment of federated telco-CDNs and analyze the potential benefit of federation to increase availability and reduce provisioning required to serve video workloads. We also revisit the potential benefits that P2P-assisted architectures provide in the light of these video access patterns. Our key findings are:

- Telco-CDN federation can reduce the provisioning cost by as much as 95%. VOD workloads benefit from federation by offloading daily peak loads and live workloads benefit by offloading unexpected high traffic triggered by regional events.
- Using P2P can lead up to 87% bandwidth savings for the CDNs during peak access hours.

Employing a strategy to filter out users who quit early by serving them using P2P can alone lead to 30% bandwidth savings for VOD traffic and 60% savings for live traffic.

Chapter Outline: We provide an overview of our dataset in Section 2.1. We analyze the implications and potential benefits for federation across telco-CDNs and for hybrid P2P-CDNs in Section 2.2 and Section 2.3 respectively before discussing related work in Section 2.4 and concluding in Section 2.5.

2.1 Dataset

The data used for this analysis was collected by `conviva.com` in real time using a client-side instrumentation library in the video player that collects information pertaining to a session. This library gets loaded when the user watches video on `conviva.com`'s affiliate content providers' websites. The library also listens to events from the player (e.g., seek, pause). The data is then aggregated and processed using Hadoop [18].

We focus on two of the most popular content providers (based in the US). These two providers appear consistently in the Top 500 sites in overall popularity ranking. Our analysis is based on data queried over two months—January 2012 and March 2012—and consists of over 30 million video viewing sessions during this period. We classify the video content into two categories:

- *VOD:* The first provider serves VOD objects that are between 35 minutes and 60 minutes long. These comprise TV series episodes, news shows, and reality show episodes.
- *Live:* The second provider serves sports events that are broadcast while the event is happening, and hence the viewing behavior is synchronized.

The VOD dataset consists of approximately 4 million users and 14 million viewing sessions and covers 1,000 video shows. The live dataset consists of around 4.5 million users and 16 million video viewing sessions covering around 10,000 different events. As in several prior studies on content popularity [51, 91], we also observe a heavy tailed Zipf distribution for overall popularity of objects for both VOD and live. While most objects have few accesses over the two months, some extremely popular objects had significant viewership. On average, users viewed 4 VOD objects and 2 live events during the course of a month, which amounts to 85 minutes of VOD objects and 65 minutes of live events per month. We also observed a few heavy hitters who watched upwards of 500 videos per month on these websites.

Session characteristics: In order to understand user behavior, we look at several characteristics of individual video sessions. Specifically, for each session we collected the following information:

- *ClientID:* The first time a client watches a video on the player, a unique identifier is assigned to the player and stored in a Flash cookie to be used by subsequent views.
- *Geographical location:* Country, state, and city of the user.
- *Provider:* Information on the AS/ISP from which the request originated.
- *Session events:* Start time and duration of the session along with details on other user interaction events like pausing and stopping.
- *Session Performance:* Average bitrate, estimated bandwidth etc. during the playback.

- *Content:* Information on the content being watched, in particular, the name of the video (which we use for classifying videos into genres) and the actual duration of the content (e.g., 45 minute show).

Region	States
1	MA, NH, TV, ME, RI, CT
2	NY, PA, NJ
3	WI, MI, IL, IN, OH
4	MO, ND, SD, NE, KS, MN, IA
5	DE, MD, DC, VA, WV, NC, SC, GA, FL
6	KY, TN, MS, AL
7	OK, TX, AR, LA
8	ID, MT, WY, NV, UT, CO, AZ, NM
9	AK, WA, OR, CA, HI

Table 2.1: List of Regions

Geographical regions: We limit our study to clients within the United States. We classified the country into 9 regions using the censor bureau designated areas [6] as shown in Table 2.1. Not surprisingly, we observe that the average load in terms of number of accesses is significantly different across different regions, and they are largely correlated with the total population of the region. We observe that this pattern holds for both live and VOD traffic except in the case of some events that have regional bias. We explore this aspect further in Section 2.2.

2.2 Analyzing Telco-CDN federation

The tremendous increase in video traffic on the Internet over the past few years has caused great challenges for ISPs. The increasing traffic has strained the ISP networks leading to higher costs and maintenance issues. However, this trend has not significantly contributed to much increase in revenue for ISPs since most of the video content is served by content providers using CDNs. As a result, several ISPs have started deploying proprietary CDNs inside their own network, providing services to content providers to serve content from caches closer to customers. This could result in increased revenue for the ISPs along with traffic reduction caused by content caching [8].

There has also been recent developments that point to interest among ISPs to deploy *telco CDN federations* by consolidating their CDN capacity and offering services to users in other ISPs [48, 70, 105]. By interconnecting telco-CDNs, consumers can reach CDN caches that are closer and are also ensured of better availability and service in case of local network congestion. Pooling resources across ISPs could potentially benefit the participating ISPs in terms of provisioning costs. It also enables ISPs to provide a global "virtual CDN" service to the content providers [8].

Although there have been pilot deployment efforts and initiatives for standardization of a federated-CDN architecture in the industry [8, 48], we are not aware of any study quantifying the benefits of telco-CDN federation, specifically in the context of Internet video. We first present video access patterns that we observed in our dataset that have significant implications for CDN

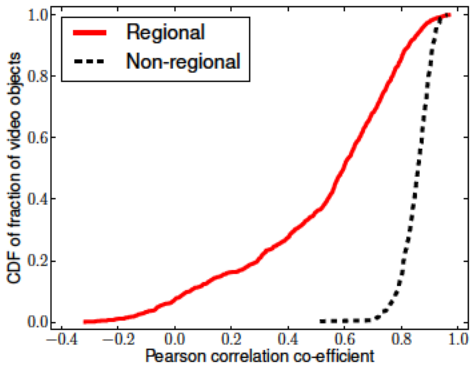


Figure 2.1: The result shows the CDF of the correlation coefficient between the #views and the population of the region for the live dataset. Non-regional content is strongly correlated with population whereas regional content is uncorrelated or negatively correlated.

federation in Section 2.2.1. We further quantify the potential benefits that telco-CDN federation can provide and put them in context with our findings on the user access patterns. To this end, we develop simple models to capture the deployment of telco-CDN federations that help us determine the potential benefits that such federation offers in Sections 2.2.2 and 2.2.3. We use this to evaluate the benefits of telco-CDN federation using our dataset for live and VOD content separately in Section 2.2.4. To the best of our knowledge this is the first large scale study to quantify the benefits of telco-CDN federation.

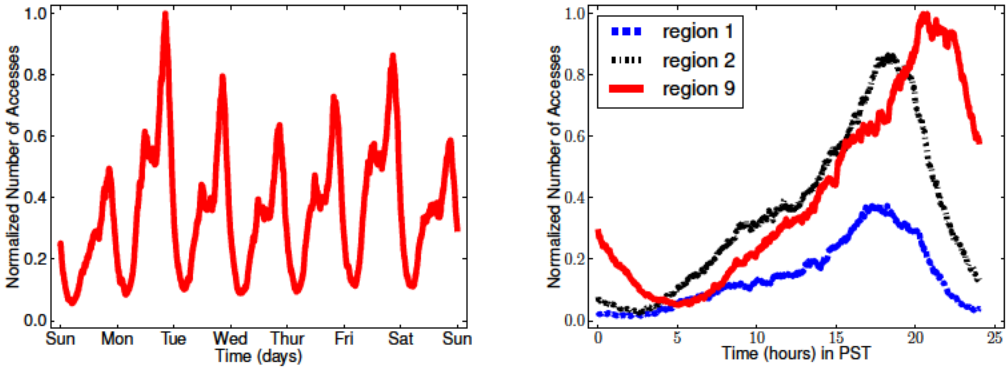
2.2.1 User Access Patterns

We observed video access patterns for live and VOD content that have implications to telco-CDN federation. For instance, in our live dataset, we observed unexpected surges in demand for certain objects from regions which can potentially be served using spare capacity in servers in other regions if CDNs federate. Similarly, we observed strong temporal shifts in when specific regions hit peak load in the VOD dataset opening up new possibilities for handling peak loads using federation. We finally also present statistics on ISP coverage and their relative performance which also have important implications when ISPs decide to federate.

Regional Interests

Typically, the number of accesses to a particular content from a geographical region is strongly correlated with the total population of the region. However, in our live dataset, we observed anomalies in the case of content with region-specific interest (e.g., when a local team is playing a game). Such unexpected surges in demands triggered by regional interests can potentially be served from servers in other regions if CDNs federate.

Our data consists of only clients within the United States and it does not contain tags with event region details. Hence, we manually classified the content as regional or non-regional based on whether it appeals to a particular region within the US. Sports matches between local teams



(a) Aggregate access pattern

(b) Region-wise access pattern

Figure 2.2: Diurnal characteristics of access pattern

within the US (e.g., NCAA) were classified as regional as opposed to events that are non-regional to the US viewers (e.g., Eurocup soccer).

We computed the Pearson correlation coefficient [101] between the number of accesses from each region to the population of the region (obtained from census data [6]). Figure 2.1 shows the CDF of the correlation coefficient across video objects for all the live objects. We observe that access rates of non-regional content show strong correlation to the population, whereas for regional matches it is uncorrelated or negatively correlated. This is because of skewed access rates from normally not so active regions because of a sporting event that has a local team. However, some regional matches show high correlation. These are highly popular events (e.g., final rounds of NCAA are of interest to everyone in the US).

Implications: The skeweness in access rates caused by regional interest is an important factor to be considered while provisioning the delivery infrastructure to handle unexpected high loads. Federation can potentially help offload such unexpected surges triggered by regional interests by using spare capacity in CDNs in other regions.

Temporal shift in peak loads

Figure 2.2a provides an overview of the VOD dataset by plotting the time series of the normalized number of videos accessed across all regions at per minute granularity for a week. As expected, we clearly observe strong time-of-day effects. To identify regional variations in peak load, we zoom into a day and plot the time series of the normalized number of accesses separately for each region in Figure 2.2b. Due to lack of space, we only show the results for the top 3 regions. The number of accesses peaks around 8 pm local time with a lull in the night. We observe that there is a difference between the time when the load peaks at different regions (caused by time zone difference). Also, we see that the peak loads are different across regions—they are largely correlated with the total population of the region.

We perform cross-correlation analysis to confirm the temporal shift in access patterns over the entire two months of data. Cross-correlation measures the degree of similarity between two time

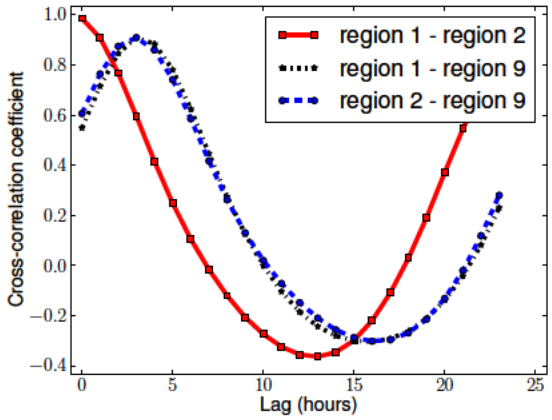


Figure 2.3: Cross correlation analysis confirms the temporal shift in access pattern over two months of data

series as a function of a time lag applied to one of them. Let $X =< X_0, X_1, ..., X_i, ... >$ denote the time series of the number of accesses as a vector where X_i is the number of accesses at time index i and $E(X_i)$ and σ_{X_i} represents the expected value and standard deviation respectively. For a given time lag k, the cross-correlation co-efficient between two time series vectors $X =< X_0, X_1, ..., X_i, ... >$ and $Y =< Y_0, Y_1, ..., Y_j, ... >$ is defined as:

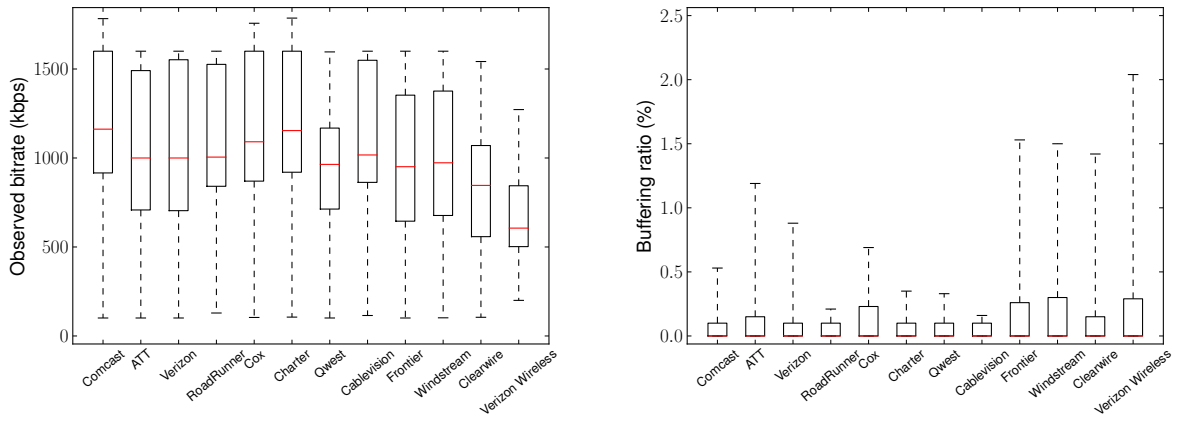
$$\tau(k) = \frac{E(X_i Y_{i+k}) - E(X_i)E(Y_{i+k})}{\sigma_{X_i}\sigma_{Y_{i+k}}} \tag{2.1}$$

The cross-correlation coefficient lies in the range of [-1,1] where $\tau(k) = 1$ implies perfect correlation at lag k and $\tau(k) = 0$ implies no correlation at lag k. We use cross-correlation to analyze the time shift in the access pattern across regions. We performed analysis across all region pairs at lags of one hour each. Due to space constraint, we present the co-coefficients plotted at different lags for the top 3 region pairs in Figure 2.3. Regions 1 and 2 fall in the same time zone and hence the $\tau(k)$ is highest at $k = 0$. Region 9 is 3 hours behind regions 1 and 2 and hence $\tau(k)$ is highest at $k = 3$. We observe this pattern holds for all the region pairs.

Implications:
The temporal shift in peak access times across different regions opens up new opportunities to handle peak loads—e.g., spare capacity at servers in regions 1 and 2 can be used to serve content in region 9 when access rates peak at region 9.

ISP performance

We study the relative performance of the ISPs over the month in terms of video quality using two key metrics identified in [72]: (1) buffering ratio defined as the percentage of session time spent in buffering, and (2) the average bitrate for each session. We summarize the relative performance of top ISPs using box-and-whiskers plots (Figure 2.4) showing the minimum, 25%ile, median, 75%ile, and 90%ile values observed across sessions. Our results corroborate a similar report



(a) Observed bitrate

(b) Buffering ratio

Figure 2.4: Performance of top ISPs

released by Netflix in May 2011 [25]. The mean performance of the ISPs are very similar, with cable ISPs like Comcast and Cox providing marginally better bitrates in the median case. We also see that wireless providers like Clearwire and Verizon Wireless provide lower bitrates compared to their wired counterparts. As observed in [72], the majority of sessions have very low buffering ratio. The median buffering ratio is zero for all the ISPs. Verizon Wireless and Windstream have marginally higher buffering ratio in the 75%ile and 90%ile case.

Implications: Since the overall performance of most ISPs are very similar, they can potentially collaboratively use their resources without worrying that their customers may see poor performance from their federating partners due to network effects.

ISP Regional presence

ISP	NY (%)	LA (%)
Comcast	1.4	1.7
AT&T	6.1	24.7
Verizon	41.7	56.3
RoadRunner	34.1	2.0
Cox	-	-
Charter	-	1.2
Qwest	-	-
Cablevision	2.9	-
Frontier	-	-
Windstream	-	-
Others	13.8	14.1

Table 2.2: Fraction of clients observed from individual ISPs for top-2 cities

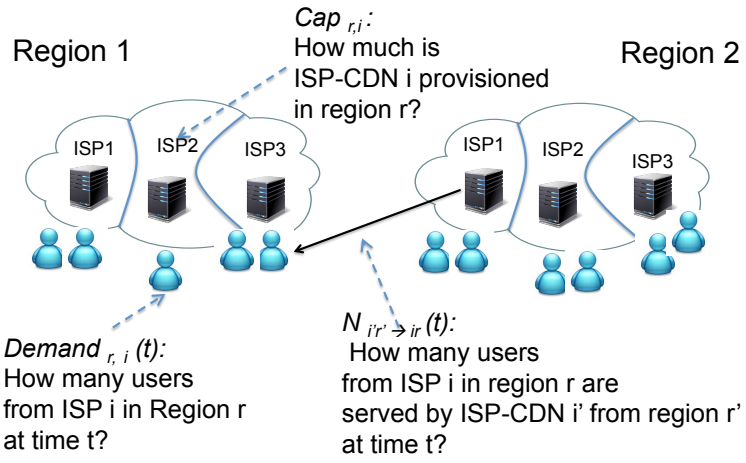


Figure 2.5: System model for telco CDN federation

Table 2.2 shows a different split across ISPs for two large cities. We observe that ISPs have significant regional biases in their coverage. For instance, while Verizon and RoadRunner have a large fraction of clients in New York city, AT&T and Verizon have a more dominant presence in LA. We also observe that some ISPs have a small fraction of their clients in cities where they are not dominant. For example, RoadRunner appears to contribute 2% of the total users in LA and AT&T has 6% in NY.

Implications: An ISP may not want to roll out new video delivery infrastructure in regions where it does not already have a significant coverage and in this case might want to direct its customers to servers located in cooperating ISPs.

2.2.2 System Model

In order to analyze the potential benefits of federation, we use a simplified system model for federated telco-CDNs. Figure 2.5 provides a high-level overview of our system model. As we have discussed earlier, there are several geographical regions, represented by $Region_r$. We currently use the regions described in Table 2.1, but this could also be more fine-grained (e.g., city or metro-area). Each region may have several ISPs, each denoted by ISP_i and each such ISP has some provisioned CDN capacity (number of users that can be served) in each region denoted by $Cap_{r,i}$.

Similar to today's ISP peering for connectivity, we envision pre-established "peering" relationships between ISPs across different regions to share their spare CDN capacity. Let $P(r,i)$ be the set of all region-ISP tuples with whom ISP_i in $Region_r$ has peering relationships. We use $r'i' \in P(r,i)$ to specify that $ISP_{i'}$ in $Region_{r'}$ has such a peering relationship with ISP_i in $Region_r$. This means that $ISP_{i'}$ in $Region_{r'}$ can offer its services or spare capacity to serve users from ISP_i in $Region_r$. This allows us to flexibly capture different kinds of telco CDN peering relationships.[1] For example, in the trivial case, without any cross-ISP federation or cross-region

[1]ISPs can also employ other relationships and policies. For example, ISPs with higher server capacity can potentially employ "provider-customer" relationships. Our current model does not capture such fine-grained policies

$$\text{Minimize: } Latency(t) + \alpha \times Dropped(t)$$

$$\forall r, i : Dropped_{r,i}(t) = Demand_{r,i}(t) - \sum_{r',i':r',i' \in P(r,i)} N_{r',i' \to r,i}(t) \tag{2.2}$$

$$\forall r, i : Dropped_{r,i}(t) \geq 0 \tag{2.3}$$

$$Dropped(t) = \sum_{r,i} Dropped_{r,i}(t) \tag{2.4}$$

$$Latency(t) = \sum_{r,i:r'i' \in P(r,i)} L_{r'i';r,i} \times N_{r',i' \to ri}(t) \tag{2.5}$$

$$\forall r', i' : \sum_{ri:r'i' \in P(r,i)} N_{r'i' \to ri}(t) \leq Cap_{r'i'} \tag{2.6}$$

Figure 2.6: Linear program for finding the optimal allocation in each logical epoch

resource sharing $P(r, i)$ relationship only contains the current ISP-region combination. In the most general case, all ISPs can share capacity with each other.

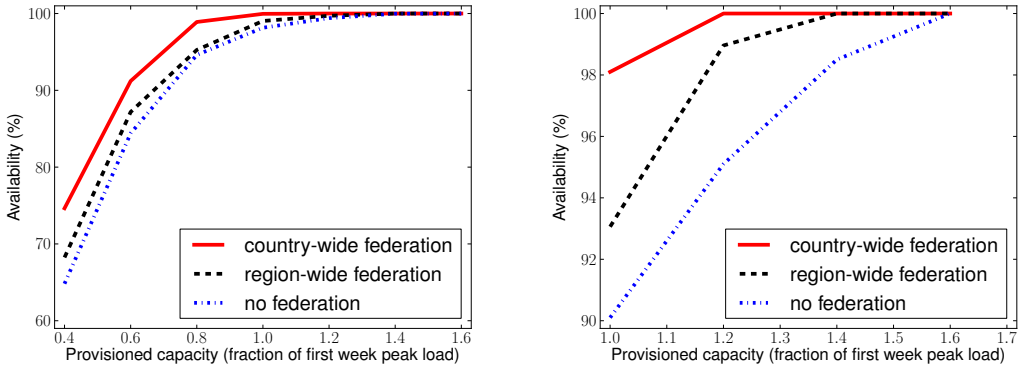
Let $Demand_{ri}(t)$ be the number of video users in region $Region_r$ from ISP_i at a given epoch t. As a first step, we only focus on the number of users and not on the specific bitrates they choose. Let $N_{r',i' \to r,i}(t)$ denote the number of users served using servers located in $ISP_{i'}$ in $Region_{r'}$ to clients from ISP_i in $Region_r$. at epoch t. We use $L_{r'i';r,i}$ to denote the latency cost incurred in this process. For clarity of discussion, we use a simple latency function at the level of "region-hops" between neighboring regions; we can extend this to incorporate more fine-grained inter-ISP latency within and across regions.

2.2.3 Global provisioning problem

Given this setup, we can now formulate the telco CDN federation problem as a *resource allocation* problem with the resources being the servers in different ISP/region combinations and the demands being the users in ISP/region combination. The linear program in Figure 2.6 formally describes the high-level optimization problem.

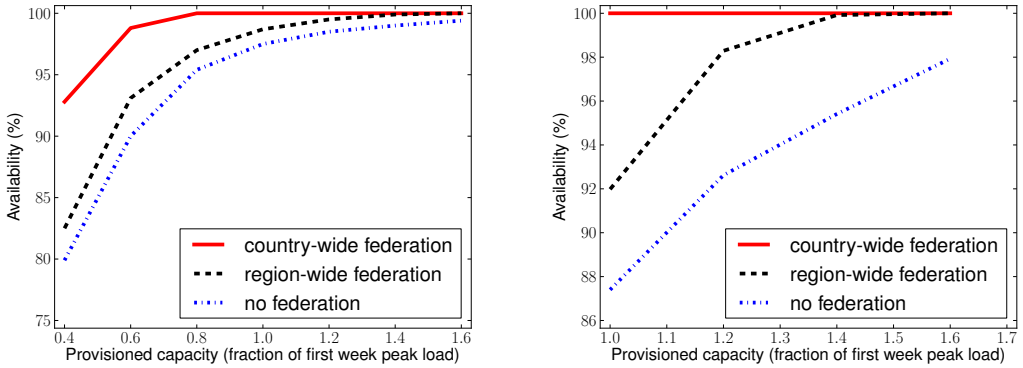
There are two high-level goals here. First, we want to accommodate as many users as possible given the current capacity provisioned at the different ISPs in various regions. Second, we want to minimize the network footprint of these assignments and ensure that requests are served as locally as possible. However, given a specific provisioning regime, it may not always be possible to fully meet the demands and some requests have to be invariably dropped. We trade off the relative importance of these objectives (i.e., latency vs. coverage) using the cost factor α in the objective function that captures the penalty for dropping users. By setting α to be very high, we can ensure that the demand is maximally met even if it requires fetching content from remote servers.

and cost models.



(a) Over multiple weeks (b) 5 hour peak access time

Figure 2.7: Benefits from federation for VOD



(a) Over multiple weeks (b) During a popular regional event

Figure 2.8: Benefits from federation for live

We capture the number of requests that are *dropped* in each ISP-region tuple Eq (2.2) and the total number of drops in Eq (2.4). (Of course, the number of requests dropped cannot be negative so we have the sanity check in Eq (2.3).) We model the overall latency footprint in Eq (2.5) using a simple weighted sum. Finally, we have a natural capacity constraint that in each region no ISP exceeds its provisioned capacity and this is captured in Eq (2.6).

2.2.4 Evaluation

We use the above formal model to evaluate the potential benefits of telco CDN federation for live and VOD content using our dataset.

Methodology: We use the user access behavior during the first week and find the peak load at each ISP_i in each $Region_r$ to determine a baseline for provisioning $Cap_{r,i}$ at each ISP-region combination. Specifically, we consider a provisioning exercise where each ISP-region combi-

nation is provisioned to handle a fraction of this peak load. Then, we use the remaining three weeks of user arrival patterns to analyze the effectiveness of such provisioning with and without federation. We set the value of α to be extremely high to minimize drops. The particular measure of interest in this exercise is the *availability* which we define as the fraction of requests that are served by the overall distribution infrastructure. Formally, this can be expressed as:

$$Availability = \frac{\sum_{r,i,t} \sum_{ri:r'i' \in P(r,i)} N_{r'i' \to ri}(t)}{\sum_{r,i,t} Demand_{ri}(t)} \tag{2.7}$$

In the following evaluation, we consider three scenarios:

- *No federation:* Here, $P(r, i)$ consists of just itself.
- *Region-wide federation:* $P(r, i)$ consists of all ISPs within the same region
- *Country-wide federation:* $P(r, i)$ consists of all ISPs in all regions.

Benefits for VOD content: Figure 2.7a shows the overall benefits of federation using the VOD dataset. As mentioned before, each telco-CDN provisions for a fraction of the observed peak load from the first week. For instance, as shown in Figure 2.7a, when each telco-CDN provisions for 40% of the observed peak load in the first week (this roughly corresponds to the average observed load), we see that there is almost a 5% increase in availability with just region-wide federation when evaluated over the workload from the next 3 weeks. Country-wide federation results in about 10% increase in availability of the system.

Although peak loads are roughly predictable for VOD content, in order to achieve 100% availability without federation, each ISP-region needs to over-provision with 1.6 times the observed first week peak load. Whereas, provisioning with 1.4 times the peak load would be enough with region-wide cooperation and provisioning with 1.2 times the observed first week peak load is sufficient to sustain the workload over the next 3 weeks with country-wide federation. This points to the fact that despite the synchrony in viewing behavior, peak loads are slightly offset across different ISPs within a region enabling using spare resources from other ISPs within the same region to improve availability. Similarly, the temporal shift in peak loads across regions due to time zone effect enables even more sharing of resources, reducing the provisioning cost to meet unexpected demands.

This result focuses on the average availability across the entire three week period. The benefits of federation are the most pronounced during peak access times. In order to highlight this further, we evaluate the availability of the system during a five-hour peak access period in Figure 2.7b. This result shows that without federation, roughly 10% of users will need to be dropped if each ISP-region was simply provisioned for the peak load observed in the first week, whereas we get only 2% dropped users with country-wide federation.

Benefits for live content: Live events have more unpredictable workloads due to interest-induced regional effects leading to unexpected higher load from typically low-load regions (e.g., when the local team plays a match). Consequently, we expect that pooling in resources from other ISPs and regions via federation will be even more beneficial.

We use the live dataset and show the overall benefits from federation in Figure 2.8a. For instance, as seen in Figure 2.8a, when provisioned for 40% of the peak load from first week, region-wide federation would increase the availability by around 3% (lower than the VOD case)

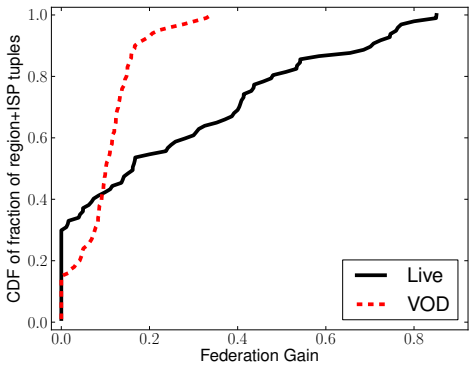


Figure 2.9: CDF of federation gain

while country-wide federation would increase the availability by 13% (higher than VOD) when evaluated on the next 3 week workload.

We zoom into a peak access time of 3 hours when a regional match was being broadcasted and repeat the study to show the benefits of federation in Figure 2.8b. We observe that employing country-wide federation, the system can achieve 100% availability by just provisioning for the observed peak load from the first week. Region-wide federation would require provisioning the system with 1.4 times the peak load. Without any federation, we observed that provisioning for 20 times the peak load is required to meet 100% availability—i.e., federation decreases the required provisioning by around 95%. This clearly shows that live events can benefit a lot from federation because unpredictable local peaks in access rates are much more common.

Which ISPs benefit the most: The immediate question that arises when we consider peering and federation is fairness. We analyze if specific categories of ISPs and/or regions are more likely to gain from federation compared to others. To this end, we define a *federation gain* metric for each ISP-region combination as the ratio between the total volume of requests served by other ISP/regions to the total capacity of this ISP-region $\frac{TotalServedbyOthers}{Capacity}$. Figure 2.9 shows the CDF of federation gain over all ISP-region combinations using country-wide federation. We observe that federation gains are lower and more uniform for VOD (highest gain is 0.4) while they are more skewed and higher in value in the case of live (highest gain is 0.8). Looking at the ISP-region combinations that benefit the most, we observe that ISPs in typically low-load regions have higher benefits in the case of live. This is because of unpredictable local peaks caused by events of regional interest. In the case of VOD, the ISPs in high-load regions have larger benefits. The benefits were mostly from offloading unexpected daily peak loads.

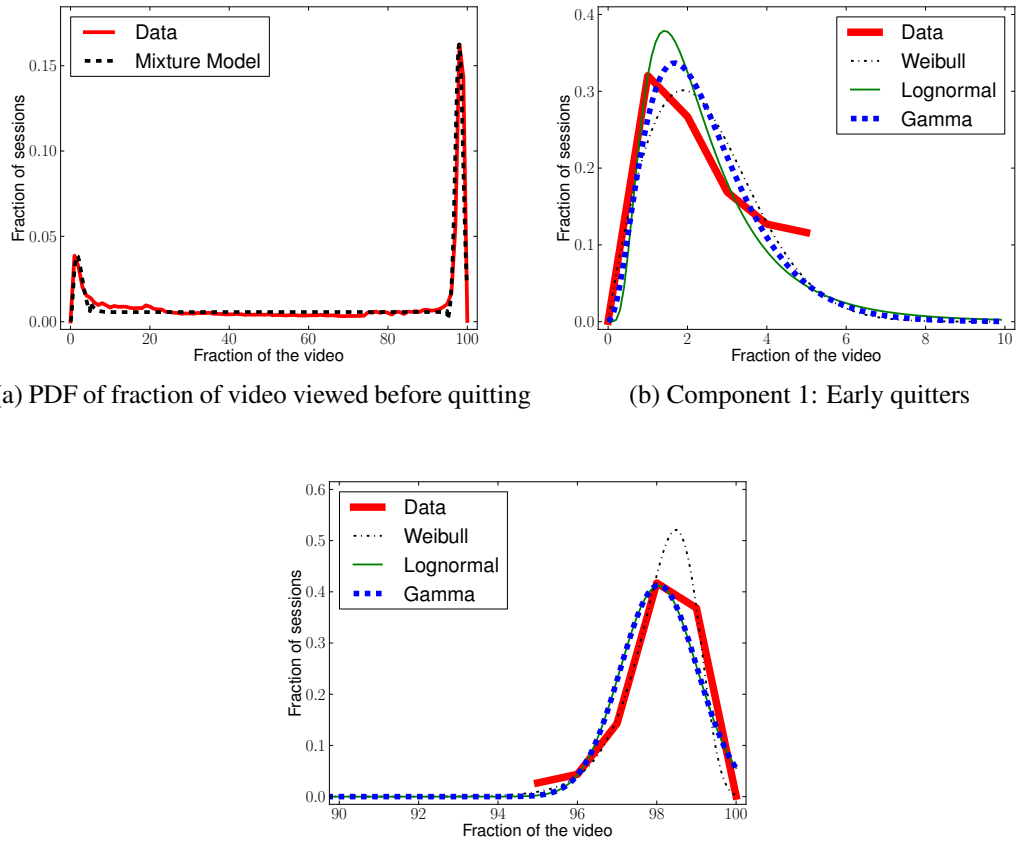
Performance Costs: Employing telco-CDN federation might lead to the selection of CDN servers far from a user, which would increase latency. Our approach to limit these performance issues is to use a very simple hop-based latency model, but a more systematic scheme would take into consideration the impact of CDN server selection on users' quality-of-experience [58, 95]. In Chapter 3, we develop models for measuring users' quality-of-experience and perform simulation studies to show initial promise on the benefits of CDN selection based on users quality-of-experience.

2.2.5 Main observations

To summarize, the key observations are:

- Federation increases the overall availability of the system with lower provisioning overhead (as much as 95% reduction in the case of live). The benefits are higher with higher level of co-operation (the upper bound being pooling in all resources within the country).
- VOD workload benefits from federation by offloading daily peak loads. We notice that ISPs from typically high load regions benefit the most.
- Live workload benefits from federation by offloading unexpected high traffic triggered by regional events. Here, the benefits are higher for ISPs in typically low-load regions.

2.3 Analyzing hybrid P2P-CDN



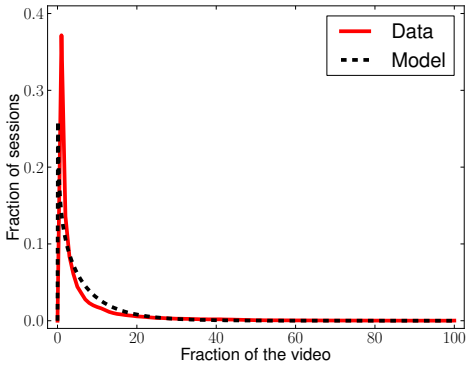
(a) PDF of fraction of video viewed before quitting

(b) Component 1: Early quitters

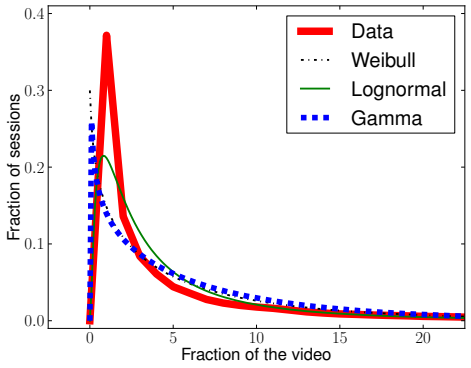
(c) Component 3: Steady viewers

Figure 2.10: Distribution of the fraction of video viewed for VOD

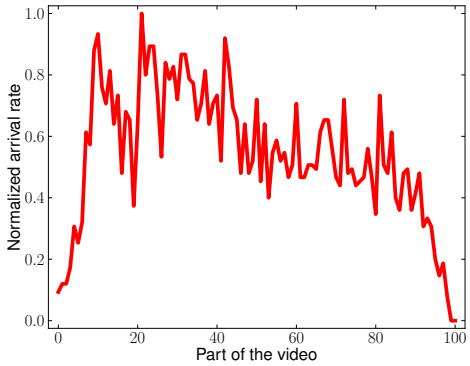
The two predominant technologies for delivering videos to end users are CDNs based on the client-server model of delivery, and server-less P2P mechanisms. While CDNs provide reliable



(a) PDF of fraction of video viewed



(b) Modeling fraction of video viewed



(c) Arrival rate

Figure 2.11: Fraction of video viewed and arrival rate for live objects

delivery using a geographically distributed delivery infrastructure, P2P enables scalability by leveraging on the bandwidth resources of individual users. There has been renewed interest in the CDN industry to augment traditional CDN based video delivery with P2P technologies. This trend is driven by the need for higher quality (e.g., [27]), and is further enabled by new technologies that allow P2P modules to run within browsers and video players without requiring separate applications [2, 4, 41].

Conventional wisdom in the use of P2P-assisted hybrid CDNs suggests that:

- P2P is only likely to be useful for live content because VOD may have low synchrony with very few users viewing the same part of the video at the same time.
- It is better to use the CDN for the early bootstrap process as clients arrive and use P2P only for the steady-state once the "swarm" dynamics stabilize.

However, we observed several user access patterns and behaviors in our dataset that give us reason to revisit and question these traditional assumptions in hybrid CDN-P2P designs. We present these observations in Section 2.3.1. Based on these observations, we propose new strategies for CDNs to reduce their infrastructure costs by using a hybrid-P2P approach and evaluate

these proposals in Section 2.3.2.

2.3.1 User Access Patterns

We observed several user access patterns that have very important implications to the design of hybrid P2P-CDN architecture. For example, we observed that several users watch only the first few minutes of a video in the case of both VOD and live content. This could imply that some parts of the video objects are more amenable to P2P than the rest. We also explore the evolution of interest for VOD and live content both within a session and also across time to understand when it would be more beneficial to employ P2P strategies.

Partial Interest in content

We observed that several users had partial interest in the content that they are viewing and they quit the session without watching the content fully in the case of both VOD and live. If most users watch only the first few minutes of the video before quitting, P2P might be more amenable for the first few chunks since there will be more copies of them compared to the rest of the video. Hence, we further investigated the temporal characteristics of user behavior within a given video session and analyzed what fraction of a video object users typically view before quitting.

For VOD content, Figure 2.10a shows that based on the fraction of video that a user viewed within a session, users can be classified into three categories:

- **Early-quitter:** A large fraction of the users watch less than 10% of the video before quitting the sessions. These users might be "sampling" the video.
- **Drop-out:** We observe that further on, users steadily drop out of the video session possibly due to quality issues or lack of interest in the content.
- **Steady viewer:** A significant fraction of the users watch the video to completion.

We can model this using a mixture model with three separate components [101]. As shown in Figures 2.10b and 2.10c, we try to find the best fitting probability distribution for the early-quitter and steady viewer components. Inspecting visually and using mean squared error test, we choose the gamma distribution to represent both the early-quitter and the steady viewer components. We model the drop-out component using a uniform distribution. We then use expectation maximization [101] to estimate the mixture model parameters and obtain the model as shown in Figure 2.10a. These models can be used for simulating video viewing behaviors in the future.

The previous result considers the behavior of users in aggregate. A natural question then is whether specific users behave in a consistent way across multiple video sessions. To this end, we profile users' viewing history across multiple sessions by grouping sessions by the user as identified using their unique *ClientID*. We find that 4.5% of the users quit the session early for more than 75% of the sessions; i.e., these users are "serial" early quitters. Similarly, 16.6% of the users consistently viewed the video to completion; i.e., these are consistently steady viewers.

Similar to the analysis that we did for VOD content, we also analyze what fraction of the live content users typically view before quitting and plot the distribution in Figure 2.11a. We observe that based on the fraction of video viewed within a session, users watching live content can be classified into two categories:

- **Early-quitter:** A very large fraction of users watch less than 20% of the video before quitting the session.
- **Drop-out:** The remaining fraction of users steadily drop out of the video session.

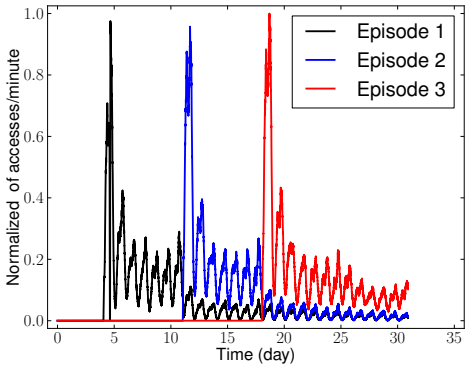
Figure 2.11b zooms into the early-quitter part of the plot and shows how well different distributions fit the data. Inspecting visually and using mean squared error test, we find that the gamma distribution is the best fit and model it in Figure 2.11a. A large fraction of users quitting the session early for live content might imply that the first part of the event is the most popular part. However, as we see in Figure 2.11c users arrive randomly within the event and stay for short periods of time before quitting. Hence the first part of the event is not necessarily the most popular part.

We also profile users' viewing history (based on the unique *Client ID*) and notice that around 20.7% of the clients are "serial" early quitters—i.e., they quit the session early for more than 75% of the sessions for live content. We also observe several users joining and quitting multiple times during the same event. Since our dataset consists of sporting events, one possibility is that they might be checking for the current score of the match.

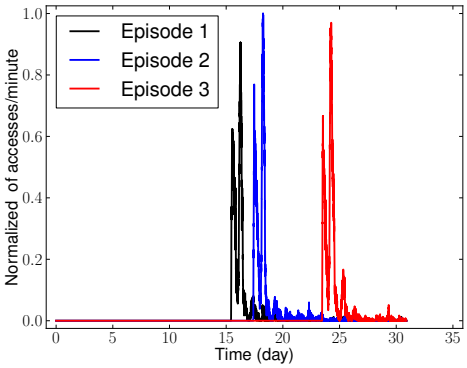
Contrasting the observations of live and VOD, we observe the following key differences:

- The early-quitters watch higher fractions of video in the case of live (up to 20% of the video) when compared to VOD (up to 10% of the video). Drop-out percentage is less pronounced in the case of live and we also do not observe a significant fraction of users viewing the entire event.
- In the case of VOD, users typically view the video from the start as opposed to live where people join at random times.
- We observe a higher fraction of "serial" early-quitters in the case of live.

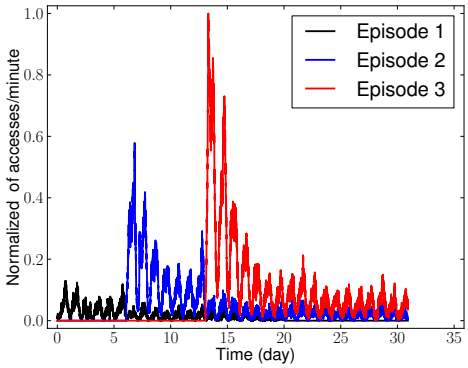
Implications:
(1) This analysis is particularly relevant in the context of augmenting CDNs with P2P based delivery. For example, if most users are likely to only watch a small fraction of video, then P2P will be less effective at offloading some of the server load as there may not be sufficient number of cached copies of the content.
(2) Content providers and content delivery infrastructures can identify the early quitters and steady viewers and customize the allocation of resources (e.g., use P2P to serve the content to early quitters who are "sampling" the video).
(3) Although user behavior like early-quitting are similar for live and VOD, we need to consider the differences in access patterns. For example, since early-quitters watch the video for longer in the case of live, employing P2P to serve early-quitters might imply serving more content using P2P in the case of live than VOD.
(4) Similarly, the fact that users typically view VOD objects from the start and quit early might imply higher availability for the first few chunks of the video. For live, even though users quit quickly, they arrive randomly in between the event and hence the first part of the event may not necessarily be the most popular part.
(5) Beyond hybrid P2P designs, this analysis is very interesting because understanding such patterns is especially useful for content providers and content delivery infrastructures in order to maximize some higher-level objective (e.g., where to place ad impressions to maximize revenue).



(a) TV series



(b) News shows



(c) Reality shows

Figure 2.12: Temporal change in popularity of VOD objects

Evolution of interest

It is crucial to investigate how popularity of content evolves over time since it could point to certain times when P2P strategies might be more beneficial. For example, if more users watch VOD videos on the day of release, there would be higher synchrony in viewership that could lead to higher benefits from employing P2P.

We classify VOD objects into three categories: TV series, news show or reality show, and model the evolution in interest along two key dimensions: (1) temporal decay in popularity for a given object (i.e., a fixed episode for a fixed show) over days, and (2) demand predictability across multiple episodes for a given show. We develop models for these parameters that can be used for simulating video workloads in the future. Live objects are viewed while the event is happening and are not available afterwards. Hence, we explore how the interest in the content evolves during the event by analyzing hotspot points in events.

Figure 2.12 shows the temporal variation in popularity and how demand for the content decays for sample objects from the three categories of VOD objects. First, for TV series episodes, the demand for episodes appears relatively stable and predictable week to week, and it decays

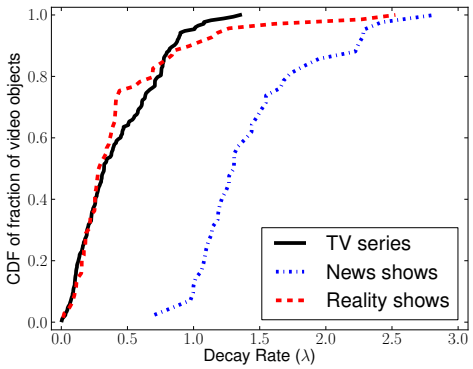


Figure 2.13: CDF of decay rate for different genres

gradually over time. Second, for news shows, we see the demand hits a peak on the release date and decreases quite dramatically. Finally, for reality shows, while we see a decay in demand from the time of release, there is less predictable viewership across different episodes. We further characterize the temporal decay and demand predictability for VOD objects.

Temporal decay in popularity for VOD objects: We observe that the highest number of accesses occurs on the day of release for all the VOD objects, and the daily peak number of access for each object decreases with time. Exponential decay appears to be the best fit for modeling the decay (compared to linear decay process) based on aggregate mean-squared error test across multiple objects. The decay in peak number of accesses can hence be characterized using an exponential decay function as follows:

$$P(t) = P_0 e^{-\lambda t} \tag{2.8}$$

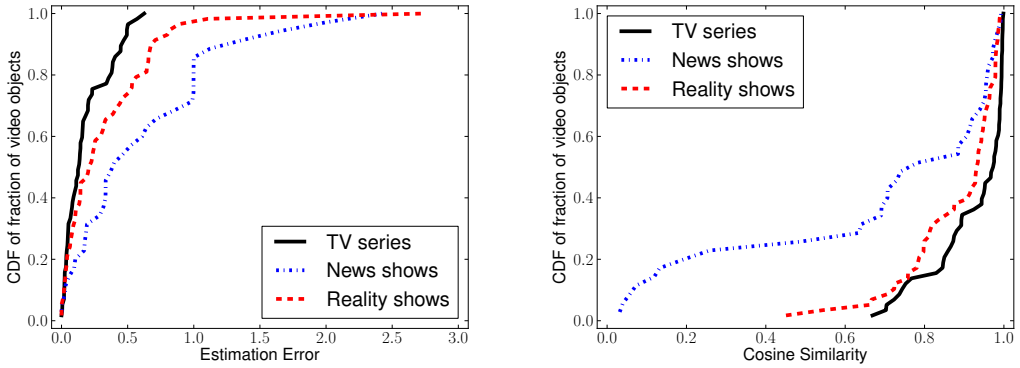
where P_0 is the peak number of access on the day of release, $P(t)$ is the peak number of access on the day t since release and λ is the decay constant. Figure 2.13 shows the CDF of the estimated decay rate (λ) for all the VOD objects categorized by their genres. News shows have high decay rates which implies that these objects turn stale quicker and their demand decreases dramatically within a day of release. In contrast, TV shows have lower decay rates. The decay rate of reality shows have more variability.

Demand predictability for VOD objects: We analyze how predictable the demand for shows are based on their viewership history. For this, we use the viewership pattern of the latest episode as an estimate for the next episode.[2] We characterize (1) how close were the peak number of accesses on the day of release? (2) how similar were the decay patterns?

(1) *Estimation Error:* Using the most recent episode as a predictor for the peak demand for the next episode, we calculate:

$$Estimation\ error = \frac{|P_{actual} - P_{estimated}|}{P_{\texttt{actual}}} \tag{2.9}$$

[2]Our dataset is limited to 2 to 4 episodes per show. Modeling viewership history over a larger span is an interesting direction for future work.



(a) Estimation Error

(b) Cosine Similarity

Figure 2.14: Characterizing demand predictability

where P_{actual} is the peak number of accesses on the day of release of the show and $P_{estimated}$ is the estimated peak number of accesses (i.e., the peak number of accesses observed for the previous show in the series). Figure 2.14a shows the CDF of relative error for different genres. We observe that TV series have lower relative error values, implying that their peak access rates across episodes are more steady and predictable. News shows and reality shows tend to have more variable peak accesses.

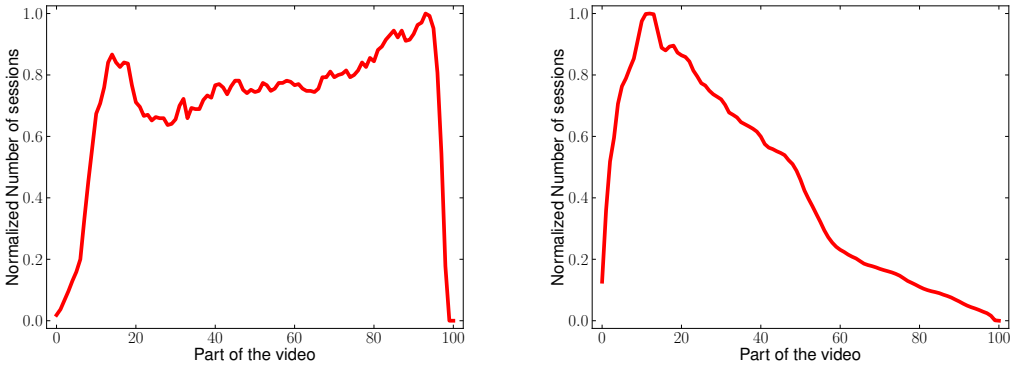
(2) *Cosine similarity:* Apart from categorizing the predictability of the peak number of accesses, we also want to estimate how similar the decay patterns are across episodes within a series. If $X =< x_0, x_1, ..., x_i, ... >$ denotes the vector of the number of accesses for the object starting from the hour of release and $Y =< y_0, y_1, ..., y_j, ... >$ denote the vector of number of accesses for the previous episode of the series, we compute the similarity between the episodes as:

$$Cosine\ similarity = \frac{\sum_{i=0}^{n} x_i \times y_i}{\sqrt{\sum_{i=0}^{n}(x_i)^2} \times \sqrt{\sum_{i=0}^{n}(y_i)^2}} \tag{2.10}$$

Cosine similarity takes values in the range [0,1] where 1 implies high similarity and 0 indicates independence.[3] Figure 2.14b shows the CDF of cosine similarity for different VOD objects. We observe that TV series have the highest similarity. The access patterns of news shows tend to be very different from the previous episodes. The cosine similarity of reality shows falls in between the TV series and news shows.

Hotspots in live events: From a provisioning perspective, it is important to understand how the interest in the content evolves during the live event. Figure 2.15 gives two extreme examples of how overall interest in the content changes within a session. Figure 2.15a shows an example of an event where the number of viewers watching the event was steady throughout the event whereas Figure 2.15b is an example of an event where there was a particular point in the event where interest peaked and then it died down. We refer to the location with the peak number of

[3]Because X and Y are both positive vectors, the cosine similarity can't be negative.



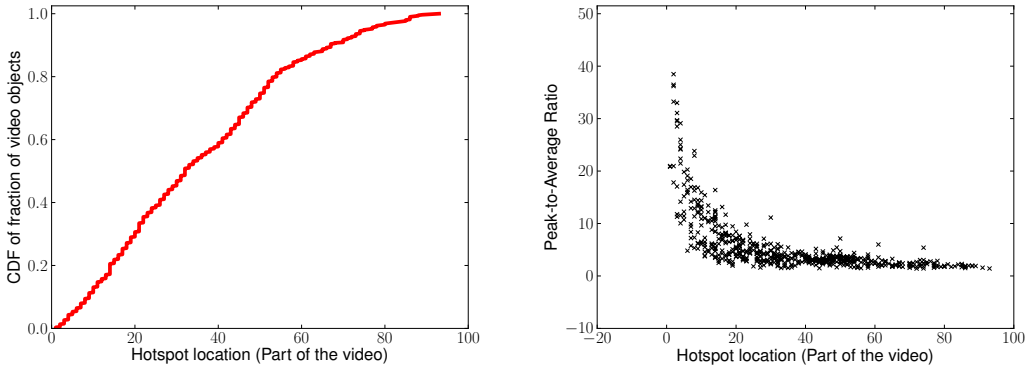
(a) Flat (b) Non-flat

Figure 2.15: Two extreme examples of temporal change in interest in live content during the duration of the event

simultaneous viewers as the *hotspot point* within the event.

Given these extremes, a natural question is what does a typical live event look like? To this end, we systematically analyze the live events on two dimensions: (1) where do hotspots occur in a video? (2) how pronounced is the hotspot? Figure 2.16a shows the CDF of the hotspot point location for all the live events. We see that there is no particular location where hotspots typically occur. To capture how pronounced a hotspot is, we compute the *peak-to-average* ratio of the number of simultaneous viewers at a given point of time during the session. Looking at the distribution of the peak-to-average ratio (Figure 2.16b), we observe that majority of the events have flat access rates (similar to Figure 2.15a). However, events with pronounced hotspots tend to have the hotspot point towards the beginning of the event.

Implications:
(1) The strong diurnal patterns observed from the time series plots again point to high synchrony of viewing even at a per-object basis. This bodes favorably in using P2P augmentation strategies for delivering VOD content.
(2) The decay rates indicate higher synchronous viewing behavior on the day of release of the show. This is also when we see higher demand in objects and when the CDNs might benefit more from using P2P strategies.
(3) Comparing genres, news shows have very high decay rates and are least predictable. This could potentially lead to sudden unexpected surges in demands and hence CDNs may need to invoke P2P-based strategies dynamically to handle these loads. However, TV series have more stable demands that are predictable and with lower decay. This means that the delivery infrastructure can be provisioned accordingly. Reality shows have much more variability in terms of decay and predictability.
(4) Since we do not observe any typical pattern for hotspot locations across live objects, CDNs may need to dynamically invoke strategies to handle the peak loads by using P2P depending on how interest evolves for the particular content.



(a) Where do hotspots occur?

(b) How pronounced are the hotspots?

Figure 2.16: Investigating hotspots in live content

2.3.2 Revisiting P2P-CDN benefits

Contrary to the conventional wisdom in this space, first, we posit that P2P might be more useful for VOD than previously assumed and that these benefits can be achieved even without assuming that each peer is caching the whole content as in [83]. Second, the presence of early quitters suggest CDNs may want to rethink how they allocate constrained server resources. Specifically, we leverage the higher interest in the early chunks coupled with the tendency of users to sample videos to consider a (possibly counter-intuitive) option where we can use P2P to bootstrap serving the content and later use the CDN servers. This allows the CDN to invest resources more usefully for viewers who are more likely to yield increased revenue from ad-impressions.

Methodology: The metric of interest here is the reduced number of accesses at the CDN server as a result of using P2P. Our goal in this exercise is to evaluate the *potential benefit* of P2P-assisted CDNs. To this end, we consider a very simplistic P2P-assisted model where peers are organized in a swarm with a specific *scope* and *size*. For each swarm, we assume that only one request needs to go to CDN server and the remaining nodes receive the content through P2P. The scope represents a trade off between network-friendliness and the availability of peers who are in synchronized viewing:

- Nodes form peers only with nodes within the *same city+ISP*.
- Based on the region classification in Table 2.1, nodes form peers only with other nodes within the *same region+ISP*.
- Nodes form peers within the *same region*

Our goal is to estimate the potential benefits of hybrid P2P-CDN. For this we further simplify the model and assume that the nodes have unlimited uplink and downlink bandwidth. We assume that the video is available at a single bitrate and all clients have sufficient bandwidth to stream at that rate. We also do not model swarm dynamics or evaluate the choice of chunk selection policies [75]. Instead, we consider a simple sequential chunk selection policy where peers are organized in swarms corresponding to the location in the video. For live content we do not consider the impact of cache size since all viewers are in sync. However, for VOD we cache

a limited number of previous chunks (e.g., several services like Netflix do not allow caching more than a few minutes of the content) and nodes typically peer with other nodes that have the required content cached. We set the chunk size to 5 minutes of the video consistent with what is predominant in the industry. We limit the maximum swarm size to 100.

Scope	Live (%)	VOD (%)
Same region	98.94	87.09
Same region+ISP	96.91	40.90
Same city+ISP	92.65	13.79

Table 2.3: Overall benefit for using P2P for different scopes

Impact of varying scope: Table 2.3 summarizes the overall benefit from using a P2P-augmented CDN system for live and VOD content. Not surprisingly, live content has higher savings than VOD. For live, the potential savings are as high as 92% even with same city+ISP scope. While the benefits are higher as we increase the scope, the resulting increase in savings shows diminishing returns. This suggests that realizing simple and network-friendly P2P solutions to augment today's CDN infrastructure is a promising path. In the case of VOD, we limit the cache size to 1 chunk. We observe that savings can be as high as 87% when nodes are allowed to peer with other nodes within the same region. Unlike live content, however, the savings are not as large when the scope of peering is limited (e.g. same city + ISP).

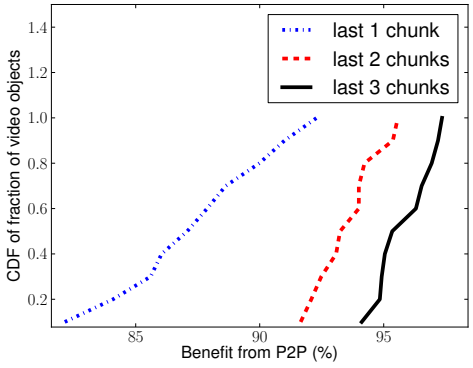


Figure 2.17: Impact of cache size on the benefit of P2P for VOD

Effect of varying cache size for VOD: In the case of VOD, it is interesting to investigate how increasing cache size affects the performance of the system. Figure 2.17 shows the CDF of the benefits from P2P for different VOD objects with same region scope. Although increasing cache size leads to greater savings, we observe diminishing returns for increased cache size.

Which part of the video gives most benefit? In Figure 2.18, we look at the percentage of savings that chunks in different parts of the video provide. We observe that most of the benefits for VOD is due to the earlier chunks. This is because users typically watch VOD videos from the start and the large number of early-quitters cause the earlier chunks to be more available than the later ones. However, in the case of live, the benefits appear to be more uniform. This

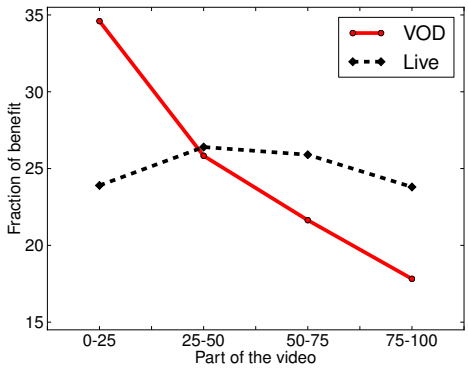


Figure 2.18: Chunks that are more likely to benefit from P2P; for VOD we see that the early chunks are the ones that benefit the most

is because of the pattern that we observed earlier—although users quit early, they also join the event at random times.

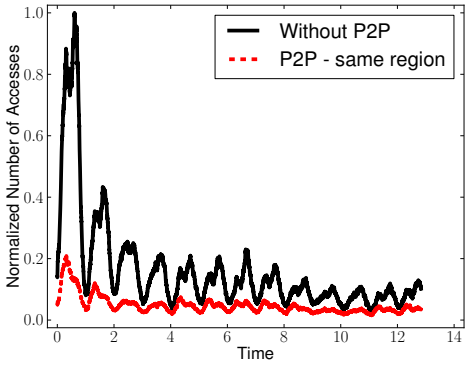


Figure 2.19: Evolution of the benefit of P2P assistance over time

How does the benefit of P2P vary over time: Unlike live events, VOD objects have demands that last several weeks. In this context, it is interesting to observe how the savings vary with time. Figure 2.19 shows the temporal variation in access demands at the server with and without P2P. We observe that the savings are as high as 80% during the peak access hour on the day of release because of larger number of users synchronously viewing content. This is also the time when the CDN would benefit the most from savings.

Using P2P earlier: Last, we explore the benefits via an alternative strategy of using P2P for the early chunks and later serving the content directly from the CDN. This can be a viewed as a mechanism to filter out the early quitters and serve them without wasting precious server resources. We analyze the benefits of serving only the first few chunks using P2P for both live and VOD in Figure 2.20. We observe that with about 2 chunks (which covers most of the early-quitter scenarios for VOD), we can get savings of around 30%. In the case of VOD, this is almost equivalent to the savings obtained from the first 2 chunks of the video (as in Figure 2.18) since

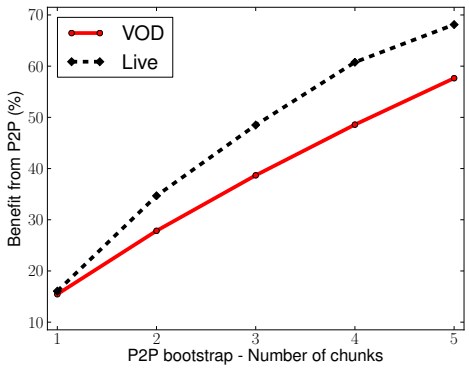


Figure 2.20: Using P2P in the early stages of user arrival

users typically watch the video from start. In the case of live, serving the first four chunks since the user starts the session (which covers all of the early-quitters) results in around 60% savings. However, note that this is not the same as the savings from the first 4 chunks of the video since users join at random times during the event and start viewing from random parts of the event.

Performance Costs: Using P2P to bootstrap video delivery might have an impact on the time it takes for the video to start playback (*start up delay*) [90]. Also, higher node churns in the P2P swarm can potentially result in disruptions in the video playback. System designs to circumvent such performance issues have been studied in previous work [118] and is not the focus of this study.

2.3.3 Main observations

To summarize, the key observations are:

- VOD has more synchrony in viewership than expected, especially during the peak access hours on the day of release. This is also when we observe the highest demand for the object. Hence, P2P can be used to offload some of the load from the server during peak hours. We observed around 87% savings for P2P-assisted VOD.
- We explore the option of bootstrapping using P2P as a means of ”filtering out” early-quitters for both VOD and live and see that this alone could lead to 30% savings in the case of VOD and 60% savings in the case of live.

2.4 Related Work

In this section, we discuss the key similarities and differences with respect to past work in measuring different aspects of Internet video.

Video performance: Previous work confirms that video quality impacts user engagement across different content genres [72, 89]. Past work also identifies that many of the quality problems observed today are a result of spatial and temporal differences in CDN performance and suggest potential workarounds via cross-CDN optimization [94, 95]. The quality problems these studies

uncover suggest that CDNs are stressed to deliver high-quality video and this motivates the need to explore strategies for augmenting CDNs.

Content popularity: There have been studies to understand content popularity in user-generated content systems (e.g., [65, 119]), IPTV systems (e.g., [51, 55, 104]), and other VOD systems (e.g., [73, 83, 91]). The focus of these studies was on understanding content popularity to enable efficient content caching and prefetching. Other studies analyze the impact of recommendation systems on program popularity (e.g., [120]) or the impact of flash-crowd like events (e.g. [117]). In contrast, our work focuses on analyzing the benefit of CDN augmentation techniques and extends these studies along two key dimensions. First, we model the longitudinal evolution in interest for different genres of video content and analyze its implications for designing a hybrid P2P-CDN infrastructure. Second, we analyze regional variations and biases in content popularity and its implications for provisioning a federated telco-CDN infrastructure.

P2P: Several pure P2P VOD systems aim to provide performance comparable to a server-side infrastructure at significantly lower cost (e.g., [56, 81, 83, 106]). There are already recent commercial efforts by CDNs to augment their infrastructures with P2P based solutions [4, 118]. Early work in the P2P space presented measurement-driven analysis on the feasibility and cost savings that hybrid-P2P technologies can bring [63, 64]. In some sense, these studies were ahead of their time—given that Internet video has really taken off only in the last 3-4 years, we believe it is critical to revisit these findings in light of new video viewing patterns. Specifically, our observations on synchronized viewing behavior for VOD and user join-leave patterns lead us to question the conventional wisdom in this space and we explore and evaluate new strategies for designing hybrid-P2P CDNs.

User behavior: Previous studies show that many users leave after a very short duration possibly due to low interest in the content (e.g., [50, 76]). While we reconfirm these observations, we also provide a systematic model for the fraction of video viewed by users using mixture model and gamma distributions, and highlight key differences between live and VOD viewing behavior. Furthermore, we analyze the implications of such partial user interest in the context of hybrid-P2P CDN deployments and explore new strategies for CDNs to reduce their bandwidth costs.

2.5 Chapter Summary

As Internet-based video consumption becomes mainstream, the video delivery infrastructure needs to be designed and provisioned to deliver high-quality content to larger user populations. But current trends indicate that the CDN infrastructure is being stressed by the increasing video traffic. Telco-CDN federation and hybrid P2P-CDNs are two oft-discussed strategies to augment existing infrastructure, but there are no recent studies on the benefits of these two strategies. Given the ongoing industry efforts and discussions to deploy federated and P2P-based solutions, we believe this work is timely: we provide a quantitative basis to justify, motivate, and inform these initiatives.

Our analysis of over 30 million live and VOD sessions reveals several interesting access patterns that have important implications to these two strategies including regional and time of day effects, synchronous viewing behavior, demand predictability, and partial interest in con-

tent. Building on these observations, we analyzed the potential benefits of hybrid P2P-CDN approaches and telco-CDN federation. We found that federation can signicantly reduce telco-CDN provisioning costs and equivalently increase the effective capacity of the system by exploiting regional and cross-ISP skews in access popularity. Surprisingly, we found that P2P approaches can work for VOD content as well, especially at peak loads when we see highly synchronous viewing patterns, and proposed and evaluated new strategies for hybrid-P2P systems based on prevalent user behavior. This study shows that even simple data analytics on user behavior data can be useful for improving content delivery.

Chapter 3

Developing a Predictive Model for Internet Video Quality-of-Experience

Video streaming forms the majority of traffic on the Internet today and its share is growing exponentially with time [9]. This growth has been driven by the confluence of low content delivery costs and the success of subscription-based and advertisement-based revenue models [10]. At the same time, users expectations for video quality are steadily rising [5]. Content providers want to maximize user engagement in order for better gains from their advertisement-based and subscription-based revenue models. Given this context, there is agreement among leading industry and academic initiatives that *improving users' quality of experience* (QoE) is crucial to sustain these revenue models [72, 87].

Despite this broad consensus, our understanding of Internet video QoE is limited. This may surprise some, especially since QoE has a very rich history in the multimedia community [23, 24, 40]. The reason is that Internet video introduces new effects with respect to both *quality* and *experience*. First, traditional quality indices (e.g., Peak Signal-to-Noise Ratio (PSNR) [29]) are now replaced by metrics that capture delivery-related effects such as rate of buffering, bitrate delivered, bitrate switching, and join time [5, 57, 72, 96, 113]. Second, traditional methods of quantifying experience through user opinion scores are replaced by new *measurable engagement measures* such as viewing time and number of visits that more directly impact content providers' business objectives [5, 113].

The goal of this chapter is to develop a *predictive model* of user QoE in viewing Internet video. To this end, we identify two key requirements that any such model should satisfy. (1) We want an **engagement-centric** model. For this, we measure user experience during a session in terms of their engagement during the session (e.g., play time). . This is based on the assumption that if users are not satisfied with the video quality during a session, they would abandon the session resulting in lower engagement metrics (e.g., lower play time). Moreover, engagement metrics such as play time can be directly tied to revenue for the content provider. For instance, trying to maximize play time would result in more advertisements streamed to the user and hence higher revenue for the content provider. (2) The model should be **actionable** and useful to guide the design of video delivery mechanisms; e.g., adaptive video player designers can use this model to tradeoff bitrate, join time, and buffering [52, 53, 74] and content providers can use it to evaluate cost-performance tradeoffs of different CDNs and bitrates [7, 96].

	Engagement-centric	Actionable
PSNR-like (e.g., [66])	✗	✓
Opinion Scores (e.g., [24])	✓	✗
Network-level (e.g., bandwidth, latency [110])	✗	✓
Single metric (e.g., bitrate, buffering)	✗	✓
Naive learning	✗	✗
Our approach	✓	✓

Table 3.1: A summary of prior models for video QoE and how they fall short of our requirements

Meeting these requirements, however, is challenging because of three key factors (Section 3.1):

- **Complex relationship between quality and engagement:** Prior measurement studies have shown complex and counter-intuitive effects in the relationship between quality metrics and engagement. For instance, one might assume that increasing bitrate should increase engagement. However, the relationship between bitrate and engagement is strangely non-monotonic [72].
- **Dependencies between quality metrics:** The metrics have subtle interdependencies and have implicit tradeoffs. For example, bitrate switching can reduce buffering. Similarly, aggressively choosing a high bitrate can increase join time and also cause more buffering.
- **Confounding factors:** There are several potential confounding factors that impact the relationship between quality and engagement: the nature of the content (e.g., live vs. Video on Demand (VOD), popularity), temporal effects (e.g., prime time vs. off-peak), and user-specific attributes (e.g., connectivity, device, user interest) [87].

As Table 3.1 shows, past approaches fail on one or more of these requirements. For instance, user opinion scores may be reflective of actual engagement, but these metrics may not be actionable because these do not directly relate to system design decisions. On the other hand, network- and encoding-related metrics are actionable but do not directly reflect the actual user engagement. Similarly, one may choose a single quality metric like buffering or bitrate, but this ignores the complex metric interdependencies and relationships of other metrics to engagement. Finally, none of the past approaches take into account the wide range of confounding factors that impact user engagement in the wild.

In order to tackle these challenges, we use large-scale data analytics, particularly machine learning to develop robust models to predict user engagement. We leverage large-scale measurements of user engagement and video session quality to run machine learning algorithms to automatically capture the complex relationships and dependencies [79]. A direct application of machine learning, however, may result in models that are not intuitive or actionable, especially because of the confounding factors. To this end, we develop a systematic framework to identify and account for these confounding factors.

Our main observations are:

- Compared to machine learning algorithms like naive Bayes and simple regression, a decision tree is more expressive to capture the complex relationships and interdependencies

and provides close to 45% accuracy in predicting engagement. Furthermore, decision trees provide an intuitive understanding into these relationships and dependencies.

- Type of video (live vs. VOD) , device (PC vs. mobile devices vs. TV) and connectivity (cable/DSL vs. wireless) are the three most important confounding factors that affect engagement. In fact, the QoE model is considerably different across different types of videos.
- Refining the decision tree model that we developed by incorporating these confounding factors can further improve the accuracy to as much as 70%.
- Using a QoE-aware delivery infrastructure that uses our proposed model to choose CDN and bitrates can lead to more than 20% improvement in overall user engagement compared to other approaches for optimizing video delivery.

Contributions and Roadmap: To summarize, our key contributions of this chapter are

- Systematically highlighting challenges in obtaining a robust video QoE model (Section 3.1);
- A roadmap for developing Internet video QoE that leverages machine learning (Section 3.2);
- A methodology for identifying and addressing the confounding factors that affect engagement (Section 3.3 and Section 3.4); and
- A practical demonstration of the utility of our QoE models to improve engagement (Section 3.5)

We discuss outstanding issues in Section 3.6 and related work in Section 3.7 before summarizing and concluding this study in Section 3.8.

3.1 Motivation and Challenges

In this section, we provide a brief background of the problem space and highlight the key challenges in developing a unified QoE model using data-driven techniques.

3.1.1 Problem scope

Multiple measurement studies have shown that video quality impacts user engagement [72, 87]. Given that engagement directly affects advertisement- and subscription-based revenue streams, there is broad consensus across the different players in the Internet video ecosystem (content providers, video player designers, third-party optimizers, CDNs) on the need to optimize video quality according to these metrics. In this study, we focus on the fraction of video that the user viewed before quitting as the measure of engagement and the following industry-standard quality metrics:

- *Average bitrate*: Video players typically switch between different bitrate streams during a single video session. Average bitrate, measured in kilobits per second, is the time average of the bitrates played during a session weighted by the time duration each bitrate was played.
- *Join time*: This represents the time it takes for the video to start playing after the user initiates a request to play the video and is measured in seconds.

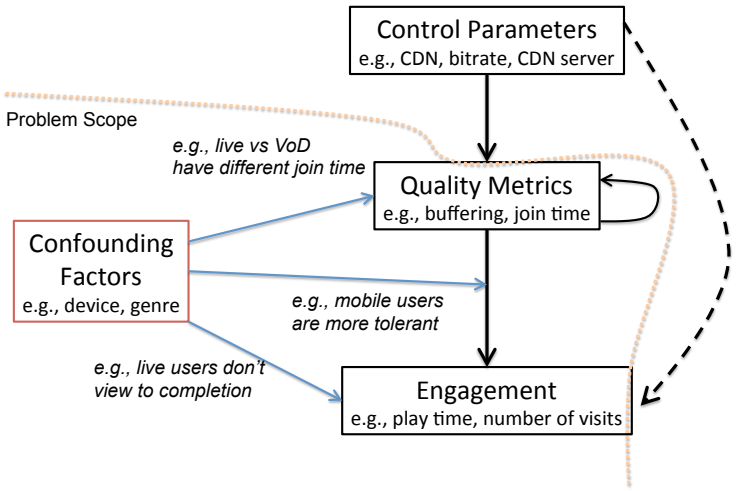


Figure 3.1: Dependencies and problem scope

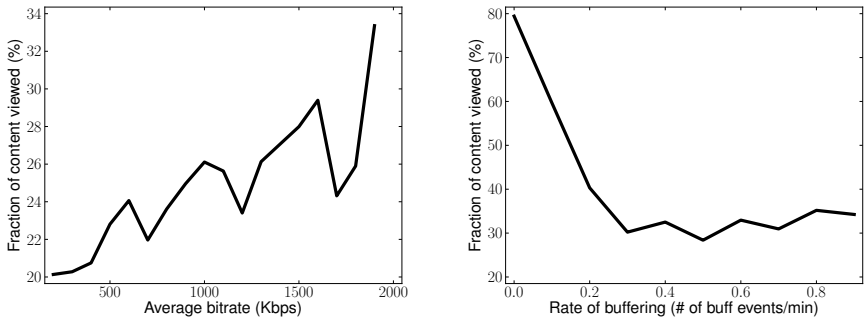
- *Buffering ratio*: It is computed as the ratio of the time the video player spends buffering to the sum of the buffering and play time and is measured in percentage.
- *Rate of buffering*: It captures the frequency at which buffering events occur during the session and is computed as the ratio of the number of buffering events to the duration of the session.

In this study, we are not reporting the impact of rate of bitrate switching due to two reasons. First, we were unable to collect this data for the large time frames that we are working on. Second, the providers use specific bitrate adaptation algorithms and we want to avoid reaching conclusions based on them. That said, the framework and the techniques that we propose are applicable if we want to later include more quality metrics to the above list.

Our goal in this study is to develop a principled *predictive model* of the user engagement as a function of these quality metrics. Such a model can then be used to inform the system design decisions that different entities in the video ecosystem make in order to tune the video according to the quality metrics to maximize engagement. For example, a video control plane may choose the CDN and bitrate for a video session based on a global optimization scheme [96], CDNs choose the servers and the specific bitrates [5], video player designers can adaptively switch bitrates [52] and so on.

An alternative direction one may consider is building a predictive model that relates the control parameters used (e.g., CDN, CDN server, adaptive bitrate selection logic) to the user engagement. While this is certainly a viable approach that a specific entity in the video delivery chain (e.g., CDN, player, "meta-CDN")[1] can employ, we believe that we can decouple this into two problems: (a) learning the relationship between the quality metrics and engagement (quality $\rightarrow$ engagement) and (b) learning the relationship between the control parameters and engagement (knobs$\rightarrow$ quality). We do so because these two relationships evolve at fundamentally different timescales and depend on diverse factors (e.g., user behavior vs. network/server artifacts). First,

[1] Some services offer CDN switching services across multiple physical CDNs to offer better cost-performance tradeoffs (e.g., [39]).



(a) Non-monotonic relationship between average bitrate and engagement

(b) Threshold effect between rate of buffering and engagement

Figure 3.2: Complex relationship between quality metrics and engagement

the control parameters available to different entities in the ecosystem may be very different; e.g., the control plane [96] operates at a coarse granularity of choosing the CDN whereas the CDN can choose a specific server. Second, the control knobs for each entity may themselves change over time; e.g., new layered codecs or more fine-grained bitrates. One can view this as a natural layering argument—decoupling the two problems allows control logics to evolve independently and helps us reuse a reference QoE model across different contexts (e.g., control plane, CDN, video player).

While modeling the knobs→quality problem is itself an interesting research challenge, this is outside the scope of this study; the focus of this study is on the problem of modeling the quality → engagement relationship.[2] As we discuss next, there are three key challenges in addressing this problem.

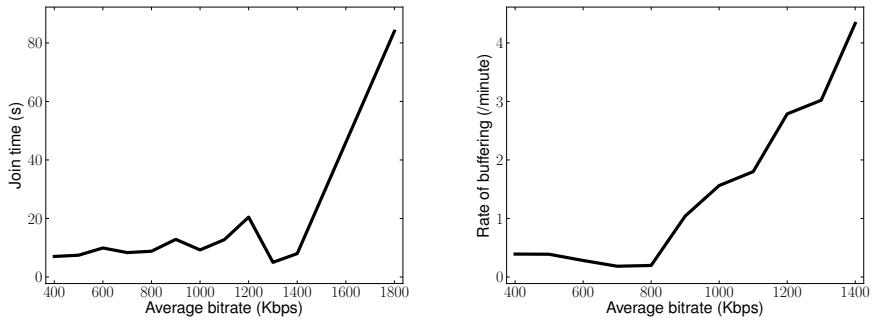
3.1.2 Dataset

The data used in this study was collected by `conviva.com` in real time using a client-side instrumentation library. This library gets loaded when users watch video on `conviva.com`'s affiliate content providers' websites. We collect all the quality metrics described earlier as well as play time for each individual session. In addition we also collect a range of user-specific (e.g., location, device, connectivity), content (e.g., live vs. VOD, popularity), and temporal attributes (e.g., hour of the day).

The dataset that is used for various analysis in this study is based on 40 million video viewing sessions collected over 3 months spanning two popular video content providers (based in the US). The first provider serves mostly VOD content that are between 35 minutes and 60 minutes long. The second provider serves sports events that are broadcast while the event is happening. Our study is limited to clients in the United States.[3]

[2]To evaluate the potential improvement due to our approach, however, we need to model this relationship as well. We use a simple quality prediction model in Section 3.5.

[3]These are distinct providers and there is no content overlap; i.e., none of the VOD videos is a replay of a live event.



(a) Higher bitrates cause higher join times (b) Higher bitrates cause higher rates of buffering

Figure 3.3: The quality metrics are interdependent on each other

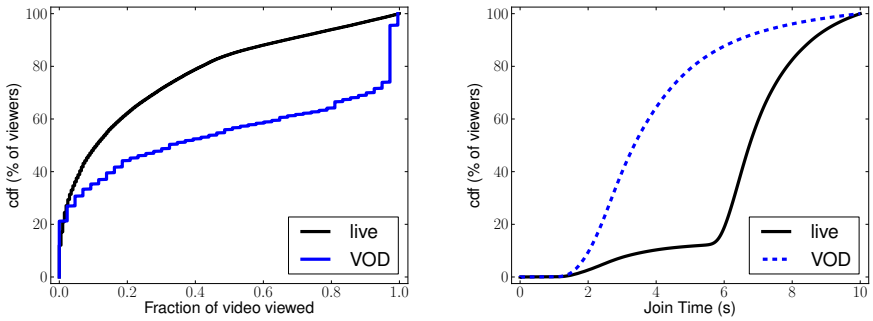
3.1.3 Challenges in developing video QoE

We use our dataset to highlight the main challenges in developing an engagement-centric of model for video QoE.

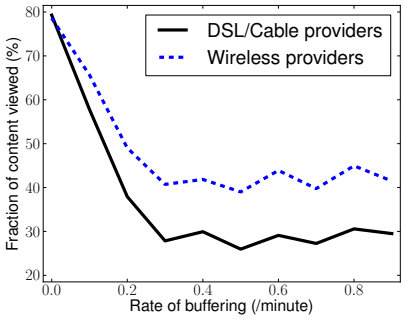
Complex relationships: The relationships between different individual quality metrics and user engagement are very complex. These were shown by Dobrian et al., and we reconfirm some of their observations [72]. For example, one might assume that higher bitrate should result in higher user engagement. Surprisingly, there is a non-monotonic relationship between them as shown in Figure 3.2a. The reason is that videos are served at specific bitrates and hence the values of average bitrates in between these standard bitrates correspond to clients that had to switch bitrates during the session. These clients likely experienced higher buffering, which led to a drop in engagement. Similarly, engagement linearly decreases with increasing rate of buffering up to a certain threshold (0.3 buffering events/minute). Beyond this, users get annoyed and they quit early as shown in Figure 3.2b.

Interaction between metrics: Naturally, the various quality metrics are interdependent on each other. For example, streaming video at a higher bitrate would lead to better quality. However, as shown in Figure 3.3a, it would take longer for the video player buffer to sufficiently fill up in order to start playback leading to higher join times. Similarly, streaming video at higher bitrates leads to higher rates of buffering as shown in Figure 3.3b.

Confounding factors: In addition to the quality metrics, several external factors also directly or indirectly affect user engagement [87]. For instance, user-attributes like user interest, content attributes like genre and temporal attributes like age of the content have effects on user engagement. A confounding factor could affect engagement and quality metrics in the following three ways (Figure 3.1). First, some factors may affect user viewing behavior itself and result in different observed engagements. For instance, Figure 3.4a shows that live and VOD video sessions have significantly different viewing patterns. While a significant fraction of the users view VOD videos to completion, live sessions are more short-lived. Second, the confounding factor can impact the quality metric. As Figure 3.4b shows, the join time distribution for live and VOD sessions are considerably different. Finally, and perhaps most importantly, the confounding factor



(a) User viewing pattern for live and VOD (b) Join time distribution for live and VOD



(c) Tolerence to rate of buffering on different connectivity

Figure 3.4: Various confounding factors directly or indirectly affect engagement

can affect the relationship between the quality metrics and engagement. For example, we see in Figure 3.4c that users on wireless connectivity are more tolerant to rate of buffering compared to users on DSL/cable connectivity.

3.2 Approach Overview

At a high-level, our goal is to express *user engagement* as a function of the quality metrics. That is, we want to capture a relationship $Engagement = f(\{QualityMetric_i\})$, where $Engagement$ can be the video playtime, number of visits to a website, and each $QualityMetric_i$ represents observed metrics such as buffering ratio, average bitrate etc. Ideally, we want this function f to be *accurate*, *intuitive*, and *actionable* in order to be adopted by content providers, video player designers, CDNs, and third-party optimization services to evaluate different provisioning and resource management tradeoffs (e.g., choosing different CDNs and bitrates).

As we saw in the motivating measurements in the previous section, developing such a model is challenging because of the complex relationships between the quality metrics and engagement, interdependencies between different quality metrics, and the presence of various confounding factors that affect the relationship between the quality metrics and engagement. In this section, we begin by presenting a high-level methodology for systematically tackling these challenges.

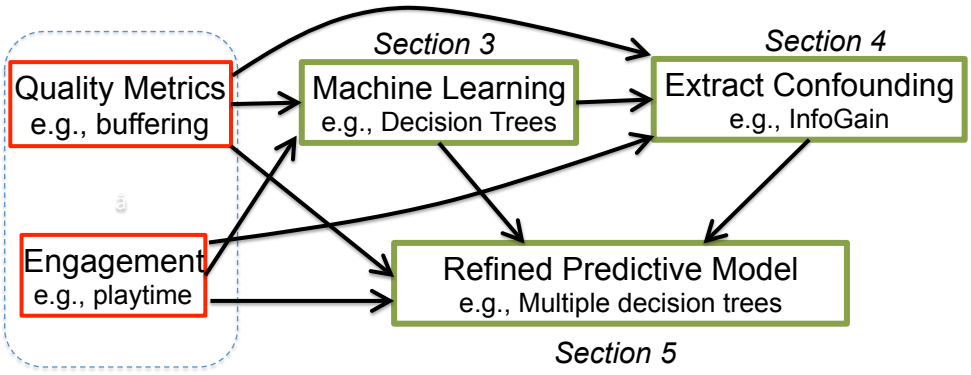


Figure 3.5: High level overview of our approach.

While the specific quality and engagement metrics of interest may change over time and the output of the prediction model may evolve as the video delivery infrastructure evolves, we believe that the data-driven and machine learning based roadmap and techniques we envision will continue to apply.

3.2.1 Roadmap

Figure 3.5 provides a high-level overview showing three main components in our approach. A key enabler for the viability of our approach is that several content providers, CDNs and third-party optimization services today collect data regarding individual video sessions using client-side instrumentation on many popular video sites. This enables a *data-driven machine learning approach* to tackle the above challenges.

Tackling complex relationships and interdependencies: We need to be careful in using machine learning as a black-box on two accounts. First, the learning algorithms must be expressive enough to tackle our challenges. For instance, naive approaches that assume that the quality metrics are independent variables or simple regression techniques that implicitly assume that the relationships between quality and engagement are linear are unlikely to work. Second, we do not want an overly complex machine learning algorithm that becomes unintuitive or unusable for practitioners. Fortunately, as we discuss in Section 3.2.2 we find that decision trees, which are generally perceived as usable intuitive models [82, 92, 114, 116] are also the most accurate. For instance, decision trees can be directly mapped into event processing rules that system designers are typically familiar with [116]. In some sense, we are exploiting the observation that given large datasets, simple non-parametric machine learning algorithms (e.g., decision trees) actually work [79].

Identifying the important confounding factors: Even though past studies have shown that external factors such as users' device and connectivity affect engagement [87], there is no systematic method to identify these factors. In Section 3.3, we propose a taxonomy for classifying potentially confounding factors. As we saw in the previous section, the confounding factors can affect our understanding in all three respects: affecting quality, affecting engagement, and also affecting how quality impacts engagement. Thus, we need a systematic methodology to identify the factors that have an impact on all three dimensions.

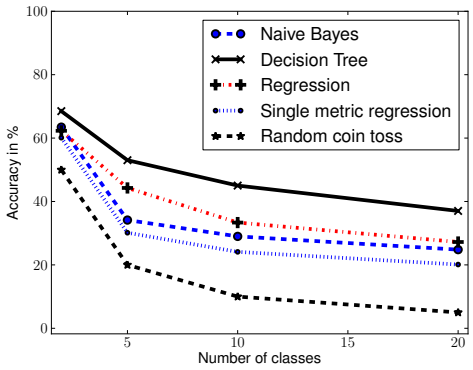


Figure 3.6: Decision tree is more expressible than naive Bayes and regression based schemes

Refinement to account for confounding factors: As we will show in Section 3.2.2, decision trees are expressive enough to capture the relationships between quality metrics and engagement. It may not, however, be expressive enough to capture the impact of all the confounding factors on engagement. In Section 3.4, we evaluate different ways by which we can incorporate these confounding factors to form a unified model.

3.2.2 Machine learning building blocks

Decision trees as predictive models: We cast the problem of modeling the relationship between the different quality metrics as a classification problem. In machine learning, classification is the problem of identifying to which of a set of categories a new observation belongs based on a model learnt from a training set consisting of data containing observations for which category membership is known. We begin by categorizing engagement into different classes based on the fraction of video that the user viewed before quitting. For example, when the number of classes is set to 5 the model tries to predict if the user viewed 0-20% or 20-40% or 40-60% or 60-80% or 80-100% of the video before quitting. We can select the granularity at which the model predicts engagement by appropriately setting the number of classes (e.g., 5 classes means 20% bins vs. 20 classes means 5% bins). We use similar domain-specific discrete classes to bin the different quality metrics. For join time, we use bins of 1 second interval; for buffering ratio we use 1% bins; for rate of buffering we use 0.1/minute bins; and for average bitrate we use 100 kbps-sized bins.

Figure 3.6 compares the performance of three different machine learning algorithms: binary decision trees, naive Bayes, and classification based on linear regression. The results are based on 10-fold cross-validation—the data is divided into 10 equally sized subsets and the model is trained 10 times, leaving out one of the subsets each time from training and tested on the omitted subset [101]. Naturally, the prediction accuracy decreases when the model has to predict at a higher granularity. We observe that decision trees perform better than naive Bayes and linear regression. This is because naive Bayes algorithm assumes that the quality metrics are independent of each other and hence it does not attempt to capture dependencies between them. Similarly, linear regression is not expressive enough to capture the complex relationships between quality

metrics and engagement. Also, as shown in Figure 3.6, performing linear regression based on just a single "best" metric (average bitrate) yields even lower accuracy since we are ignoring the complex metric interdependencies and the relationships between other metrics and engagement.

Information gain analysis: *Information gain* is a standard approach for uncovering hidden relationships between variables. More importantly, it does so without making any assumption about the nature of these relationships (e.g., monotone, linear effects); it merely identifies that there is some potential relationship. Information gain is a standard technique used in machine learning for feature extraction—i.e., identifying the key features that we need to use in a prediction task. Thus, it can serve as a natural starting point for systematically identifying confounding factors.

The information gain is based on the idea of entropy of a random variable Y which is defined as $H(Y) = \sum_i P[Y = y_i] log \frac{1}{P[Y=y_i]}$ where $P[Y = y_i]$ is the probability that $Y = y_i$. It represents the number of bits that would have to be transmitted to identify Y from n equally likely possibilities. The lesser the entropy the more uniform the distribution is. The conditional entropy of Y given another random variable X is $H(Y|X) = \sum_j P[X = x_j]H(Y|X = x_j)$. It represents the number of bits that would be required to be transmitted to identify Y given that both the sender and the receiver know the corresponding value of X. Information gain is defined as $H(Y) - H(Y|X)$ and it is the number of bits saved on average when we transmit Y and both sender and receiver know X. The relative information gain can then be defined as $\frac{H(Y)-H(Y|X)}{H(Y)}$.

In the next section, we use the information gain analysis to reason if a confounding factor impacts either engagement or quality.

Compacted decision trees: Decision trees help in categorizing and generalizing the data given in the dataset and provide a visual model representing various if-then rules. One main drawback while dealing with multi-dimensional large datasets is that these techniques produce too many rules making it difficult to understand and use the discovered rules with just manual inspection or other analysis techniques [92]. In order to get a high-level intuitive understanding of the impact of different quality metrics on engagement, we compact the decision tree. First, we group the quality metrics into more coarse-grained bins. For instance, we classify average bitrate into three classes—very low, low, and high. The other quality metrics (buffering ratio, buffering rate, and join time) and engagement are classified as either high or low. We then run the decision tree algorithm and compact the resulting structure to form general rules using the technique described in [92]. The high-level idea is to prune the nodes whose majority classes are significant; e.g., if more than 75% of the data points that follow a particular rule belong to a particular engagement class then we prune the tree at that level. The tree formed using this technique may not be highly accurate. Note that the goal of compacting the decision tree is only to get a high-level understanding of what quality metrics affect engagement the most and form simple rules of how they impact engagement. Our predictive model uses the original (i.e., uncompressed) decision tree; we do not sacrifice any expressive power. In the next section, we use this technique to test if a confounding factor impacts the relationship between quality metrics and engagement—particularly to check if it changes the relative importance of the quality metrics.

3.2.3 Limitations

We acknowledge three potential limitations in our study that could apply more broadly to video QoE measurement.

- **Fraction of video viewed as a metric for engagement:** While fraction of video viewed before quitting may translate into revenue associated from actual advertisement impressions, it does not capture various psychological factors that affect engagement (e.g., user may not be interested in the video and might be playing the video in the background). We use fraction of video viewed as a measure of engagement since it can be easily and objectively measured and it provides a concrete starting point. The high-level framework that we propose can be applied to other notions of engagement.
- **Coverage over confounding factors:** There might be several confounding factors that affect engagement that are not captured in our dataset (e.g., user interest in the content). Our model provides the baseline in terms of accuracy—uncovering other confounding factors and incorporating them into the model will lead to better models and higher prediction accuracy.
- **Early quitters:** A large fraction of users quit the session after watching the video for a short duration. These users might be either "sampling" the video or quitting the session because of quality issues. They can be treated in three ways: (1) Remove them completely from the analysis, (2) Separate them into two groups based on their quality metrics (high quality population and low quality population) and learn a separate model for each group (3) Profile users based on their viewing history and predict whether they will quit early or not based on their interest in the video content. We use (1) in this thesis as it provides a clearer understanding of how quality impacts engagement. That said, approaches (2) and (3) are likely to be useful and complementary in a system-design context; e.g., to guide resource-driven tradeoffs on which users to prioritize.

3.3 Identifying Confounding Factors

In this section, we propose a framework for identifying confounding factors. To this end, we begin with a taxonomy of potentially confounding factors. Then, we use the machine learning building blocks described in the previous section to identify the factors that have a non-trivial impact on engagement.

3.3.1 Approach

We identify three categories of potential confounding factors from our dataset:

- **Content attributes:** This includes the *type of video* (i.e., live vs. VOD) and the *overall popularity* (i.e., number of views).
- **User attributes:** This includes the user's *location* (region within continental US), *device* (e.g., smartphones, tablets, PC, TV), and *connectivity* (e.g., DSL, cable, mobile or wireless).

- **Temporal attributes:** Unlike live content that is viewed during the event, VOD objects in the dataset are available to be accessed at any point in time since its release. This opens up various temporal attributes that can possibly affect engagement including the *time of day* and *day of week* of the session and the *time since release* for the object (e.g., day of release vs. not).

We acknowledge that this list is only illustrative as we are only accounting for factors that can be measured directly and objectively. For example, the user's interest in the particular content is also a potential confounding factor that we cannot directly measure. Our goal here is to develop a systematic methodology to identify and account for these factors. Given more fine-grained instrumentation to measure other types of factors (e.g., use of gaze tracking in HCI), we can use our framework to evaluate these other factors as well.

In Section 3.1, we saw that confounding factors can act in three possible ways:

1. They can affect the observed engagement (e.g., Figure 3.4a)
2. They can affect the observed quality metric and thus indirectly impact engagement (e.g., Figure 3.4b);
3. They can impact the nature and magnitude of quality $\rightarrow$ engagement relationship (e.g., Figure 3.4c).

For (1) and (2) we use *information gain analysis* to identify if there is a hidden relationship between the potential confounding factor w.r.t engagement or the quality metrics. For (3), we identify two sub-effects: the impact of the confounding factor on the quality $\rightarrow$ engagement relationship can be *qualitative* (i.e., the relative importance of the different quality metrics may change) or it can be *quantitative* (i.e., the tolerance to one or more of the quality metrics might be different). For the qualitative effect, we use the compacted decision tree separately for each class (e.g., TV vs. mobile vs. PC) using the method described in Section 3.2.2 and compare their structure. Finally, for the quantitative sub-effect in (3), we simply check if there is any significant difference in tolerance.

3.3.2 Analysis results

Next, for each user, content, and temporal attribute we run the different identification techniques to check if it needs to be flagged as a potential confounding factor. Table 3.2 presents the information gain between each factor with respect to engagement (fraction of video viewed before quitting) and the four quality metrics.

Type of video: Classified as live or VOD session, as Table 3.2 shows, the type of video has high information gain with respect to engagement confirming our earlier observation that the viewing behavior for live and VOD are different (Section 3.1.3). Again, since join time distributions for live and VOD sessions are also different (Section 3.1.3), it is not surprising that we observe high information gain in join time. Similarly, the set of bitrates used by the live provider and the VOD provider are quite different leading to high information gain for average bitrate as well.

We learn the compacted decision tree for VOD and live sessions separately as shown in Figure 3.7 and see that the trees are structurally different. While buffering ratio and rate of buffering have the highest impact for VOD, average bitrate has the highest impact for live events. Somewhat surprisingly, some live users tolerate very low bitrates. We believe this is related to

Confounding Factor	Engagement	Join Time	Buff. Ratio	Rate of buff.	Avg. bitrate
Type of video (live or VOD)	8.8	15.2	0.7	0.3	6.9
Overall popularity (live)	0.1	0.0	0.0	0.2	0.4
Overall popularity (VOD)	0.1	0.2	0.4	0.1	0.2
Time since release (VOD)	0.1	0.1	0.1	0.0	0.2
Time of day (VOD)	0.2	0.6	2.2	0.5	0.4
Day of week (VOD)	0.1	0.2	1.1	0.2	0.1
Device (live)	1.3	1.3	1.1	1.2	2.7
Device (VOD)	0.5	11.8	1.5	1.5	10.3
Region (live)	0.6	0.7	1.3	0.5	0.4
Region (VOD)	0.1	0.3	1.2	0.2	0.2
Connectivity (live)	0.7	1.1	1.4	1.1	1.5
Connectivity (VOD)	0.1	0.4	1.1	1.4	1.3

Table 3.2: Relative information gain (%) between different potential confounding factors and the engagement and quality metrics. We mark any factor with more than 5% information gain as a potential confounding factor

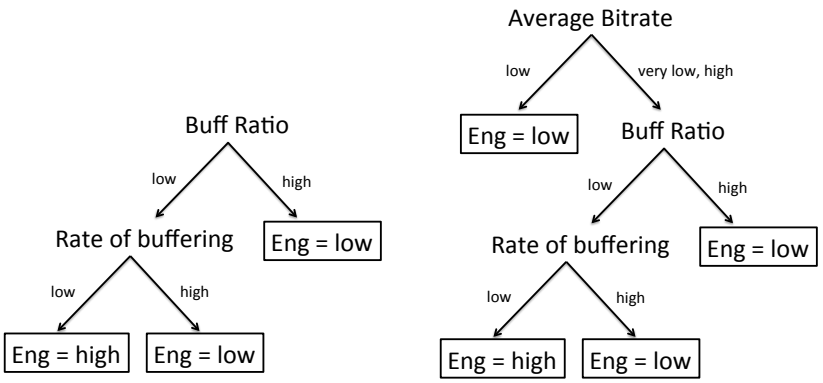
an observation from prior work which showed that users viewing live sporting events may run the video in background and the player automatically switches to lower quality to reduce CPU consumption [72].

Since the differences between live and VOD sessions are considerably large, for the remaining attributes, we perform the analysis separately for live and VOD data.

Device: We classify the devices as PC (desktops and laptops) or mobile devices (smartphones and tablets) or TV (e.g., via Xbox). In the VOD dataset, 66% of the traffic were initiated from PC and around 33% were from TV. Mobile users formed a small fraction. However, in the live dataset, 80% of the traffic were initiated from PCs and almost 20% of the traffic from mobile users—users on TV formed a small fraction.

For a subset of the VOD data, we observed that the compacted decision tree for the TV users was different from that of mobile and PC users. While PC and mobile users showed a tree structure similar to Figure 3.7a, we observed Figure 3.9 in the case of TV. Intrigued by this difference, we visualized the impact of bitrate on engagement. Curiously, we find in Figure 3.8 that increased bitrate led to lower engagement in the case of TV. This is especially surprising as one would expect that users would prefer higher bitrates on larger screens. Investigating this further, we saw complaints on the specific content provider's forum regarding contrast issues at higher bitrate when viewing the video on TV. This issue was later corrected by the provider and the newer data did not show this anomaly. As shown in Table 3.2, we observe a high information gain in terms of join time and average bitrate for VOD data.

Even though the compacted tree was similar in structure for TV, PC and mobile users (not shown), Figure 3.10 and 3.11 show substantial differences in tolerance levels for average bitrate and rate of buffering. This is consistent with a recent measurement study that shows that mobile users are more tolerant toward low quality [87]. The one difference with live data, however, is that device does not lead to high information gain for engagement or any of the quality met-



(a) VOD (b) Live

Figure 3.7: Compacted decision tree for live and VOD are considerably different in structure

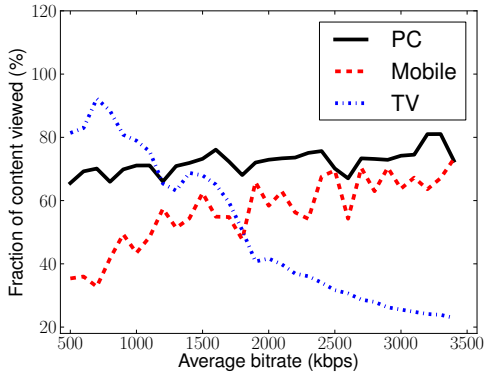


Figure 3.8: Anomalous trend : Higher bitrate led to lower engagement in the case of TV in the VOD dataset

rics (Table 3.2). Because of the differences in tolerance, we consider device as an important confounding factor.

Connectivity: Based on the origin ISP, we classify the video session as originating from a DSL/cable provider or from a wireless provider. In the VOD dataset, 95% of the sessions were initiated from DSL/cable providers. In the live dataset, 90% were from DSL/cable providers. We see that sessions with wireless connection had slightly lower bitrates and higher buffering values compared to the sessions in cable and DSL connection. This accounts for the slight information gain that we observe in Table 3.2.

The compacted decision trees had the same structure for cable vs. wireless providers for both live and VOD data. But we observed difference in tolerance to rate of buffering for both live and VOD content. As we observed earlier in Section 3.1.3, users were more tolerant to buffering rate when they were on a wireless provider for VOD content. For live content, as shown in Figure 3.12, we observed difference in tolerance for rate of buffering. Due to these differences, we consider connectivity as a confounding factor.

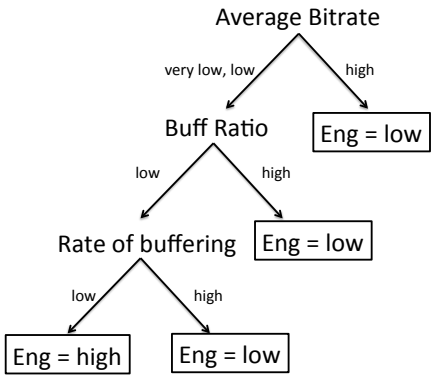
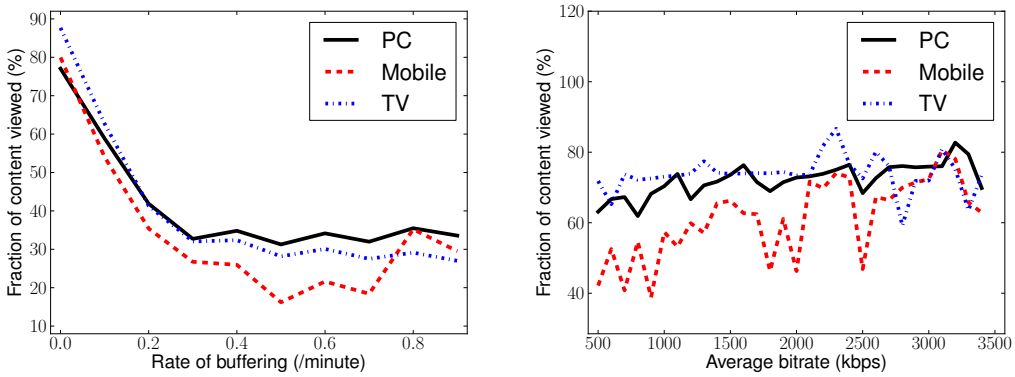


Figure 3.9: Compacted decision tree for TV for the VOD data that showed the anomalous trend



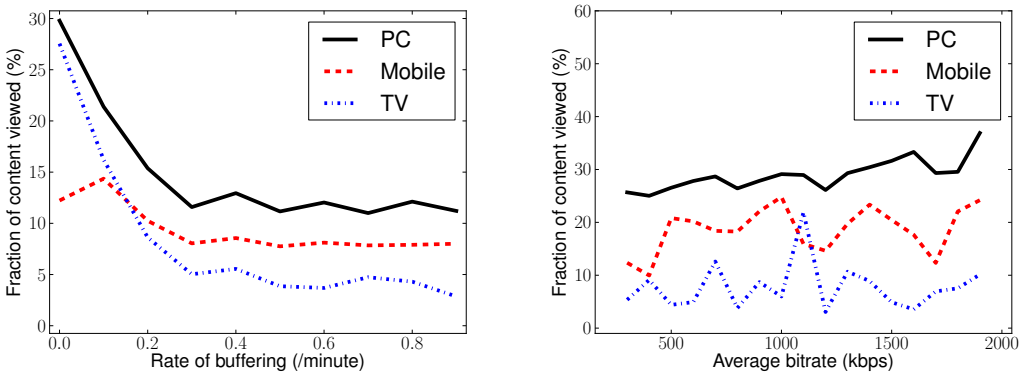
(a) Rate of buffering

(b) Average bitrate

Figure 3.10: VOD users on different devices have different levels of tolerance for rate of buffering and average bitrate

Time of day: Based on the time of the day, the sessions were classified as during night time (midnight-9am), day time (9am-6pm) or peak hours(6pm-midnight). We observed that 15% of the traffic were during night time, 30% during day time and 55% during peak hours. We also observed that users experienced slightly more buffering during peak hours when compared to late nights and day time. The compacted decision trees were similar for peak hours vs. day vs. night. Users were, however, slightly more tolerant to rate of buffering during peak hours as shown in Figure 3.13. Since we want to take a conservative stance while shortlisting confounding factors, we consider time of day to be a confounding factor.

Other factors: We did not observe high information gain or significant differences in the compacted decision tree or the tolerance to quality for other factors such as region, popularity, day of week, and time since video release (not shown). Thus, we do not consider these as confounding factors.



(a) Rate of buffering (b) Average bitrate

Figure 3.11: Live users on different devices have different levels of tolerance for rate of buffering and average bitrate

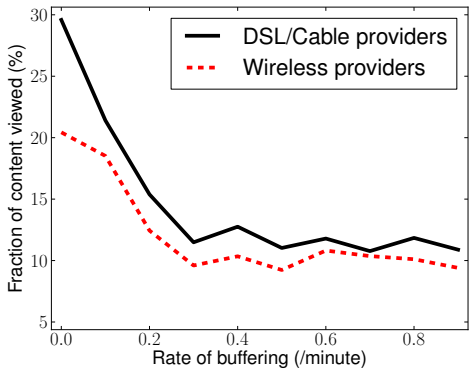


Figure 3.12: For live content, users on DSL/cable connection and users on wireless connection showed difference in tolerance for rate of buffering

3.3.3 Summary of main observations

Table 3.3 summarizes our findings from the analysis of various potential confounding factors. Our main findings are:

- The main confounding factors are type of video, device, and connectivity.
- The four techniques that we proposed for detecting confounding factors are *complementary* and expose different facets of the confounding factors.
- Our model also reconfirmed prior observations on player-specific optimizations for background video sessions. It was also able to reveal interesting anomalies due to specific player bugs.

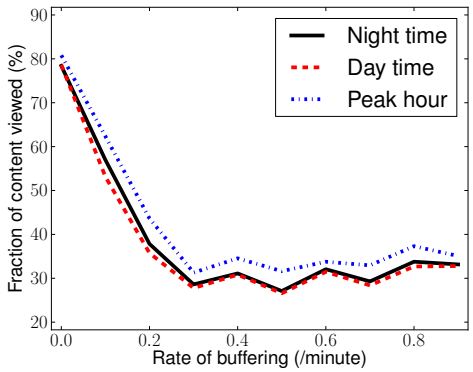


Figure 3.13: For VOD, users tolerance for rate of buffering is slightly higher during peak hours

Confounding Factor	Engmnt	Quality	Q→E Qual	Q→E Quant
Type of video - live or VOD	✓	✓	✓	✓
Overall popularity (live)	✗	✗	✗	✗
Overall popularity (VOD)	✗	✗	✗	✗
Time since release (VOD)	✗	✗	✗	✗
Time of day (VOD)	✗	✗	✗	✓
Day of week (VOD)	✗	✗	✗	✗
Device (live)	✗	✗	✗	✓
Device (VOD)	✗	✓	✓ ✗	✓
Region (live)	✗	✗	✗	✗
Region (VOD)	✗	✗	✗	✗
Connectivity (live)	✗	✗	✗	✓
Connectivity (VOD)	✗	✗	✗	✓

Table 3.3: Summary of the confounding factors. Check mark indicates if a factor impacts quality or engagement or the quality→engagement relationship. The highlighted rows show the key confounding factors that we identify and use for refining our predictive model

3.4 Addressing confounding factors

Next, we describe how we refine the basic decision tree model we saw in Section 3.2.2 to take into account the key confounding factors from the previous section. We begin by describing two candidate approaches for model refinement and the tradeoffs involved. We study the impact of both candidate approaches and choose a heuristic "splitting" based approach.

3.4.1 Candidate approaches

There are two candidate approaches to incorporate the confounding factors into the predictive model:

- **Add as new feature:** The simplest approach is to add the key confounding factors as additional features in the input to the machine learning algorithm and relearn the prediction

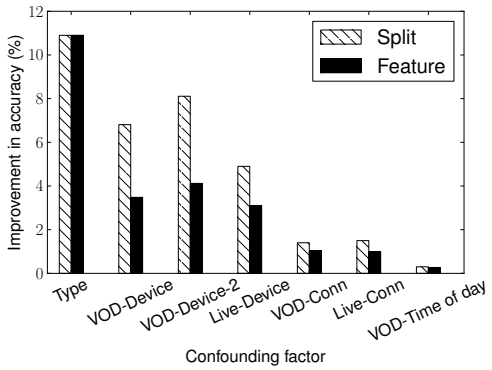


Figure 3.14: Comparing feature vs split approach for the different confounding factors

model.

- **Split Data:** Another possibility is to split the data based on the confounding factors (e.g., live on mobile device) and learn separate models for each split. Our predictive model would then be the logical union of multiple decision trees—one for each combination of the values of various confounding factors.

Both approaches have pros and cons. The feature-addition approach has the appeal of being simple and requiring minimal modifications to the machine learning framework. This assumes that the learning algorithm is robust enough to capture the effects caused by the confounding factors. Furthermore, it will learn a single unified model over all the data. The augmented model we learn, however, might be less intuitive and less amenable to compact representations. Specifically, in the context of the decision tree, mixing quality metrics with confounding factors may result in different levels of the tree branching out on different types of variables. This makes it harder to visually reason about the implications for system design. For instance, consider the scenario where we want to know the impact of quality on engagement for mobile users in order to design a new mobile-specific bitrate adaptation algorithm for live sports content. This is a natural and pertinent question that a practitioner would face. To answer this question, we would in effect have to create a new "projection" of the tree that loses the original structure of the decision tree. Moreover, we would have to create this projection for every such system design question.

In contrast, the split data approach will explicitly generate these intuitive projections for different combinations of the confounding factors by construction. It also avoids any doubts we may have about the expressiveness of the machine learning algorithm. The challenge with the split approach is the "curse of dimensionality". As we have more factors to split, the available data per split becomes progressively sparser. Consequently, the model learned may not have sufficient data samples to create a robust enough model.[4] Fortunately, we have two reasons to be hopeful in our setting. First, we have already pruned the set of possibly confounding external factors to the key confounding factors. Second, as Internet video traffic grows, we will have larger datasets to run these algorithms and that will alleviate concerns with limited data for

[4]Note that this dimensionality problem is not unique to the split data approach. A decision tree (or any learning algorithm for that matter) faces the same problem as we go deeper into the tree as well. The split approach just elevates the dimensionality problem to the first stage of the learning itself.

multiple splits.

Following in the data-driven spirit of our approach, we analyze the improvements in prediction accuracy that each approach gives before choosing one of these techniques.

3.4.2 Results

For this study, we set the number of classes for engagement to 10. We observe similar results for other number of classes as well. Figure 3.14 compares the increase in accuracy using the feature and the split approach for the three key confounding factors.

As shown in Figure 3.14, splitting based on type of video vs. adding it as a feature (*Type*) results in the same increase in accuracy. In the case of the split approach, we observe that both splits (live and VOD) do not have the same accuracy—live is more predictable than VOD. However, splitting based on the device type gives better improvement compared to adding device as a feature for both VOD and live (*VOD-Device*, *VOD-Device-2* and *Live-Device*). But, we observed that the accuracy across the splits were not the same. For the VOD dataset, splits corresponding to TV and PC had higher accuracy compared to the split corresponding to smartphones. This is because, as we saw in Section 3.3, only a small fraction of users viewed VOD content on mobile phones in our dataset. *VOD-Device-2* corresponds to the data in which we observed an anomalous trend in viewing behavior on TV. Here, we observed that the split corresponding to TV had very high accuracy leading to better gains from splitting. For the live dataset, we however observed that the TV split had lower gains compared to mobile and smartphones. This is again because of the inadequate amount of data—the fraction of users watching live content on TV in our dataset was negligible.

Splitting works better than feature addition for both live (*Live-Conn*) and VOD (*VOD-Conn*) in the case of connectivity and for time of day in the case of VOD (*VOD-Time of day*). Time of day did not lead to a huge improvement in improvement in accuracy and hence we ignore it. The other external factors that we considered in Section 3.3 led to negligible increase in accuracy when addressed using both these approaches.

Why does split perform better? Across the various confounding factors, we see that the split data approach is better (or equivalent) to the feature addition approach. The reason for this is related to the decision tree algorithm. Decision trees use information gain for identifying the best attribute to branch on. Information gain based schemes, however, are biased towards attributes that have multiple levels [71]. While we bin all the quality metrics at an extremely fine level, the confounding factors have only few categories (e.g., TV or PC/laptop or smartphone/tablet in the case of devices). This biases the decision tree towards always selecting the quality metrics to be more important. In the case of type of video, the information gain in engagement is very high since user viewing behavior is very different (i.e, it satisfies criteria number (1) that we have for identifying confounding factors). So it gets chosen at the top level and hence splitting and adding as a feature led to same gain.

3.4.3 Proposed predictive model

As mentioned in Section 3.2.3, we observed many users who "sample" the video and quit early if it is not of interest [119]. Taking into account this *domain-specific observation*, we ignore

Model	Accuracy (in %)
Simple decision tree	45.02
Without early-quitters	51.20
Multiple decision tree	68.74

Table 3.4: Summary of the model refinements and resultant accuracy when number of classes for engagement is 10

these early-quitter sessions from our dataset and relearn the model leading to $\approx 6\%$ increase in accuracy.

Further incorporating the three key confounding factors (type of device, device and connectivity), we propose a unified QoE model based on splitting the dataset for various confounding factors and learning multiple decision trees—one for each split. Accounting for all the confounding factors further leads to around 18% improvement. Table 3.4 summarizes the overall accuracies when number of classes for engagement is set to 10. This implies that about 70% of the predictions are within the same 10% bucket as the actual user viewing duration.

3.5 Implications for system design

In this section, we demonstrate the practical utility of the QoE model using trace-driven simulations. We simulate a video control plane setting similar to previous work and use our QoE model to guide the choice of CDN and bitrate [96]. We compare the potential improvement in engagement using our QoE model against other strawman solutions.

3.5.1 Overview of a video control plane

The QoE model that we developed can be used by various principals in the Internet video ecosystem to guide system design decisions. For instance, video player designers can use the model to guide the design of efficient bitrate adaptation algorithms. Similarly, CDNs can optimize overall engagement by assigning bitrates for each individual client using our QoE model.

Prior work makes the case for a coordinated control plane for Internet video based on their observation that a purely client- or server- driven bitrate and CDN adaptation scheme for video sessions might be suboptimal [96]. This (hypothetical) control plane design uses a global view of the network and CDN performance to choose the CDN and bitrates for each session based on a global optimization framework. The goal of the optimization is to pick the right control parameters (in this case, CDN and bitrate) in order to maximize the overall engagement. As shown in Figure 3.15, this control plane takes as input control parameters (CDN, bitrate) and other attributes (device, region, ISP etc.) as input and predicts the expected engagement.

Our QoE model can be used to guide the design of such a video control plane. Note, however, that the QoE model from Section 3.4 takes various quality metrics and attributes that are confounding as input and predicts the expected engagement. Thus, as discussed in Section 3.1, we also need to develop a quality model which takes CDN, bitrate, and client attributes as input and predicts the quality metrics (buffering ratio, rate of buffering and join time) in order to fully

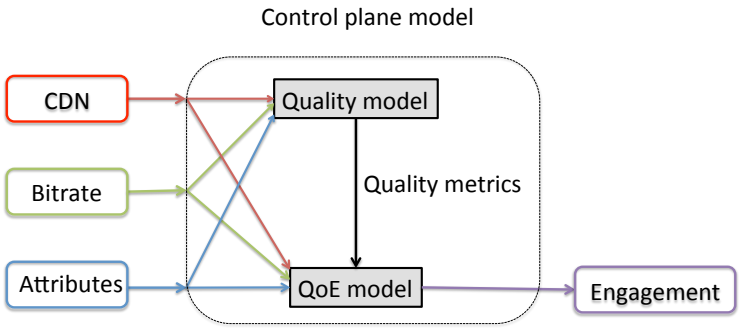


Figure 3.15: We use a simple quality model along with our QoE model to simulate a control plane. The inputs and outputs to the various components are shown above.

realize this control plane design. Figure 3.15 shows the various components and their inputs and outputs. Our key contribution here is in demonstrating the use of the QoE model within this control plane framework and showing that a QoE-aware delivery infrastructure could further improve the overall engagement. Developing a model to accurately predict the quality metrics is out of the scope of this chapter. However, it is an interesting line for future work.

3.5.2 Quality model

We use a simplified version of the quality prediction model proposed from prior work [96]. It computes the mean performance (buffering ratio, rate of buffering and join time) for each combination of attributes (e.g., type of video, ISP, region, device) and control parameters (e.g., bitrate and CDN) using empirical estimation. For example, we estimate the performance of all Comcast clients in the east coast of the United States that streamed live content over an Xbox from Akamai at 2500 Kbps by computing the empirical mean for each of the quality metrics.

When queried with specific attributes (CDN and bitrate) the models returns the estimated performance. One challenge, however, is that adding more attributes to model often leads to data sparsity. In this case, we use a hierarchical estimation heuristic—i.e, if we do not have sufficient data to compute the mean performance value for a specific attribute, CDN and bitrate combination, we use a coarser-grained granularity of attribute elements [96]. For example, if we do not have enough data regarding the performance of Xbox over Akamai over Comcast connection from the east coast at 2500 Kbps, we return the mean performance that we observed over all the devices over Akamai at 2500 Kbps over Comcast connection from the east coast. We follow the following hierarchy for this estimation: $\{$Type of video, ISP, region, device$\}$ $<$ $\{$Type of video, ISP, region$\}$ $<$ $\{$Type of video, ISP$\}$.

3.5.3 Strategies

We compare the following strategies to pick the control parameters (CDN and bitrate):

1. Smart QoE approach: For our smart QoE approach, we use a *predicted quality model* and a *predicted QoE model* based on historical data. For choosing the best control parameters for a particular session, we employ the following brute force approach. We estimate the expected engagement for all possible combinations of CDNs and bitrates by querying the *predicted quality model* and the *predicted QoE model* with the appropriate attributes (ISP, device etc.). This approach assigns the CDN, bitrate combination that gives the best predicted engagement.

2. Smart CDN approaches: We find the best CDN for a given combination of attributes (region, ISP and device) using the *predicted quality model* by comparing the mean performance of each CDN in terms of buffering ratio across all bitrates and assign clients to this CDN. We implement three variants for picking the bitrate:

2(a) Smart CDN, highest bitrate: The client always chooses to stream at the highest bitrate that is available.

2(b) Smart CDN, lowest buffering ratio: The client is assigned the bitrate that is expected to cause the lowest buffering ratio based on the *predicted quality model*

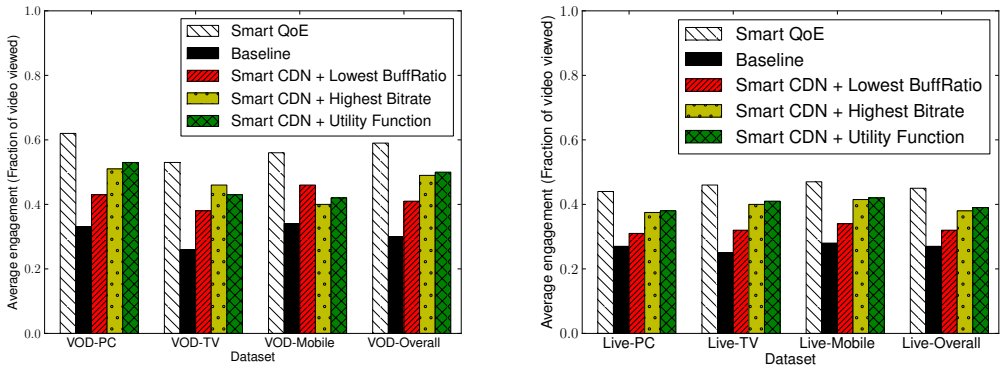
2(c) Smart CDN, control plane utility function: The client is assigned the bitrate that would maximize the utility function ($-3.7 \times BuffRatio + \frac{Bitrate}{20}$) [96].

3. Baseline: We implemented a naive approach where the client picks a CDN and bitrate randomly.

3.5.4 Evaluation

To quantitatively evaluate the benefit of using the QoE model, we perform a trace based simulation. We use week-long trace to simulate client attributes and arrival times. In each epoch (one hour time slots), a number of clients with varying attributes (type of video, ISP, device) arrive. For each client session, we assign the CDN and bitrate based on the various strategies mentioned earlier. For simplicity, we assume the CDNs are sufficiently provisioned and do not degrade their performance throughout our simulation. To evaluate the performance of these strategies, we develop *actual engagement models* and an *actual quality models* based on the empirical data from the current measurement epoch and compare the engagement predicted by these models for each session. (Note that the models that we use for prediction are based on historical data). Since the arrival patterns and the client attributes are the same for all the strategies, they have the same denominator in each epoch.

Figure 3.16 compares the performance of the different strategies for live and VOD datasets broken down by performance on each device type. We compared the different approaches under the same simulation scenario and hence do not show the error bars in these plots. As expected, the baseline scheme has the worst performance. The smart QoE approach can potentially improve user engagement by up to $2\times$ compared to the baseline scheme. We observed that the smart CDN and lowest buffering ratio scheme picks the lowest bitrates and hence the expected engagements are lower compared to the other smart schemes (except in the case of VOD on mobile phones where it outperforms the other smart CDN approaches). The smart CDN with utility function approach and smart CDN highest bitrate approaches have very comparable performances. This



(a) VOD (b) Live

Figure 3.16: Comparing the predicted average engagement for the different strategies

is because the utility function favors the highest bitrate in most cases. Our smart QoE approach picks intermediate bitrates and dynamically shifts between picking the lower and the higher bitrates based on the various attributes and the predicted quality. Thus, it can potentially improve user engagement by more than 20% compared to the other strategies.

3.6 Discussion

Other engagement measures: Content providers are also interested in other measures of engagement involving different time scales of user involvement such as customer return probability to the particular service. The quality metrics might impact these other engagement measures differently [72]; e.g., join time may affect the return probability of the customer even though it does not have a huge impact on the user engagement during a particular session. We may have to weigh in these different notions of engagement to compute an aggregate engagement index. Having said that, we believe the high-level data-driven approach we propose can be applied to other notions of engagement as well.

Hidden confounding factors and personalization: In this study, we identified a few confounding factors based on features that were captured in our dataset. There could be several other confounding factors that affect engagement that are not captured in our dataset. One obvious example is users' interest in the content. Similarly, there could potentially be inherent user preferences to quality—for instance, while Person A might prefer higher bitrate, person B might prefer lower number of bitrate switches. Identifying such individual preferences or better cohorts of users with similar preferences and incorporating them into the model would result in better accuracy and as a result lead to better adaptive streaming algorithms. In fact, user preferences for quality metrics can be considered as yet another hidden confounding factor.

Evolution of the QoE model: As the network and the user expectations for quality change with time, the QoE model also needs to evolve to capture these effects. For instance, the specific bitrates at which providers serve content might change with time. Similarly, with time, users

might have higher expectations with respect to quality from the video delivery system. In this context, we envision a live refinement system that constantly observes and analyzes the user viewing habits and continuously adapts the QoE model based on these observations.

QoE model for other Internet services: The methodology that we proposed can be generalized to be used for developing QoE models for other Internet services as well. The specific metrics, confounding factors and inferences might be different, but the general methodology of developing a data-driven predictive QoE model using machine learning techniques can be applied to new domains like web browsing, VoIP, online gaming etc. In the next chapter, we develop a model for mobile web browsing quality of experience using a similar approach.

3.7 Related Work

Engagement in Internet video: Past measurement studies have shown that video quality impacts user engagement [72, 87]. However, they provide a simple quantitative understanding of the impact of individual quality metrics (e.g., buffering) on engagement. We shed further light and provide a unified understanding of how all the quality metrics when put together impact engagement by developing a QoE model. Similarly, previous studies have also shown that a few external factors (e.g., connectivity) affect user engagement [87]. Recent work suggests the use of Quasi Experimental Design (QED) to eliminate any possible bias that can be caused by confounding factors and establish causal relationships [87]. However, these factors have to be provided a priori and there does not exist any techniques to identify if an external factor is potentially confounding or not. We extend our previous work [57] by developing techniques to identify external factors that are confounding and incorporate these factors to form a unified QoE model.

User studies: Prior work by the multimedia community try to assess video quality by performing subjective user studies and validating objective video quality models against the user study scores [40, 68, 86, 99, 103]. User studies are typically done at a small-scale with a few hundred users and the perceptual scores given by users under a controlled setting may not translate into measures of user engagement in the wild. The data-driven approach that we proposed is scalable and it produces an engagement-centric model.

Control plane: Liu et al., make a case for a co-ordinated video control plane that uses measurement-driven parameters to improve video quality by adapting CDN and bitrate of clients using a global optimization scheme [96]. As we showed, our QoE model can be used under a similar framework to further improve the benefits.

Adaptive video players design: Commercial video players today perform client-side bitrate adaptation based on current bandwidth conditions [35]. Studies that have analyzed these players have found that there is significant scope for improving their adaptation schemes [52]. Video player designers typically use ad hoc mechanisms to trade-off between various network parameters [52, 53]. Our video QoE model can be potentially used to develop engagement-centric video player adaptation algorithms.

Diagnosis: Past work has looked at techniques to proactively diagnose quality issues in video delivery in order to minimize its impact on users [98, 115]. In Section 3.3.2, we show that

our model can also detect anomalous behavior among users watching VOD content on TV, and potentially other quality issues as well.

QoE metrics in other media: There have been attempts to study the impact of network factors on user engagement and user satisfaction in the context of other media technologies. For example, in [49], the authors study the impact of bitrate, jitter and delay on call duration in Skype and propose a unified user satisfaction metric as a combination of these factors. Our approach derives a unified QoE model in the context of Internet video and it is very timely given that Internet video has gone mainstream in the past few years.

Other video measurement studies: Several measurement studies of Internet video have focused on content popularity, user behavior and viewing patterns [76, 119]. The observations made in these works have implications on understanding measurement-driven insights and performing domain-specific refinements to improve the QoE model. For instance, Yu et al., also observed that users sample videos and quit the session early [119]. Similarly, we observed that some users tolerate low bitrate while watching live content. Previous work also observed this phenomena which is a result of the player running in the background [72].

3.8 Chapter Summary

An imminent challenge that the Internet video ecosystem—content providers, content delivery networks, analytics services, video player designers, and users—face is the lack of a reference methodology to measure the *Quality-of-Experience* (QoE) that different solutions provide. With the "coming of age" of this technology and the establishment of industry standard groups (e.g., [12]), such a measure will become a fundamental requirement to promote further innovation by allowing us to objectively compare different competing designs [37, 54].

Internet video presents both a challenge and an opportunity for QoE measurement. On one hand, the nature of the delivery infrastructure introduces new complex relationships between quality and engagement and between quality metrics themselves. To further make matters worse, there are many confounding factors introduced by different aspects of this ecosystem that directly or indirectly impact engagement (e.g., genre, popularity, device). At the same time, however, we have an unprecedented opportunity to obtain a systematic understanding of QoE because of the ability to collect large client- and server-side measurements of actual user behavior *in the wild*.

This study is a significant first step in seizing this opportunity and addressing the above challenges. We developed a data-driven machine learning approach to capture the complex interactions as well as confounding effects. We also demonstrated significant practical benefits that content providers can obtain by using an improved QoE prediction model over current ad hoc approaches.

Chapter 4

Predictive Analytics for Extracting and Monitoring Web Performance over Cellular Networks

Mobile web data usage is predicted to increase eleven-fold between 2013 and 2018 [9], and web browsing is already one of the most dominant applications on cellular networks [108]. However, cellular networks are not as well designed for the web as their wireline counterparts and hence they are slower [3]. Increasing user expectations for quality is posing challenges to cellular network operators to configure their networks better by adding the right infrastructure for better Quality of Experience (QoE). In this chapter, we build predictive models for web browsing QoE. Similar to the study in Chapter 3, we use engagement metrics (e.g., number of clicks) as a proxy for quality of experience based on the assumption that if a user is not satisfied with the quality of the session, it would result in them quitting the session early resulting in lower engagement metrics (e.g., lower number of clicks). Moreover, engagement metrics are easier to measure/estimate compared to more subjective quality of experience metrics such as Mean Opinion Score (MOS). We use a set of web user experience metrics, such as session length (the number of pages a user clicks through) and abandonment (whether a user leaves a website after visiting the landing page) for our study, which we hereafter refer to as *web QoE*.

The ability to monitor web QoE is essential to determining when and where degraded network conditions actually impair user experience. Moreover, understanding the relationship between web QoE and radio factors can help troubleshoot such conditions and help operators evaluate the inherent trade-offs in potential solutions. For example, an operator may want to decide whether to increase a cell's transmit power to improve signal strength, or decrease it to reduce handovers from overlapping cells.

Prior work on monitoring web QoE relies heavily on client-side or server-side instrumentation such as browser plugins and server logs. Past work has studied the impact of web page complexity on user experience [62, 84], developing better browsers [100], detecting inefficiencies in HTTP [102] etc. These works have led to best practices that have helped improve website designs, browsers, and network protocols. However, the network between the website and the user also plays a critical role in the user experience, particularly in wireless mobile networks. To complement these studies, we take a "cellular operator view" of web QoE. Understanding

web QoE from an operator's perspective is more challenging because, unlike other players such as content providers and CDNs, network operators do not have access to detailed client-side or server-side logs, any feedback from the end hosts, or any a priori domain knowledge about website structure. Hence, it is imperative for network operators to accurately estimate QoE metrics using only network measurements.

To complicate matters, websites have evolved from serving relatively static objects, such as hypertext and images, to hosting rich mobile media applications. These sites typically deliver dynamic and personalized content that often includes third-party content such as advertising [62]. Such design typically involves fetching large number of objects from multiple domains and servers. This significant change in web page structure and content over the last decade makes accurate estimation of web QoE metrics from mobile network traces even more challenging. A key challenge is that there is not yet a scalable and accurate method to distinguish between different mobile browsing sessions or to associate each HTTP transaction with a browsing session, by only observing flow-level network traffic. Previous approaches (e.g., [84]) were designed for the "desktop" web and fare poorly when applied to mobile websites.

To the best of our knowledge, this study presents the first large-scale measurement-driven study that characterizes and models mobile web QoE and relates it to the measurable radio network characteristics, such as radio signal strength, handovers, data rate, etc. To this end, we use a month-long anonymized data set collected from a Tier-1 US-based cellular network and analyze web browsing sessions of 3 leading mobile websites that consistently appear in the top 100 [31]. For this analysis, we design mechanisms to measure and evaluate the two key web QoE metrics that quantify user experience using only network data: session length (number of pages a user clicks through) and abandonment rate (if a user leaves the website after visting the landing page). Moreover, we show that partial download ratio (fraction of page objects that download incomplete) is a measure that likely captures user disatisfaction even for single-click sessions (the majority for some websites), as it correlates well with the other two metrics.

We make the following contributions:

- We design and evaluate a novel technique to reconstruct mobile web sessions and detect user clicks[1] from HTTP traces, and demonstrate that it significantly outperforms current state-of-the-art method. Our approach is based on bag-of-words and Naive Bayes [101], an approach borrowed from text classification, and extracts features from HTTP headers. It detects clicks on mobile websites with about 20% higher recall and higher precision compared to the previous state-of-the-art [84].
- We quantify the individual impact of various network characteristics on mobile web QoE in the wild, and derive actionable findings for cellular operators. For example, we find that web QoE is very sensitive to inter-radio-access-technology (IRAT) handovers: most sessions with IRAT handovers were abandoned. Somewhat surprisingly, we find that web QoE is not noticeably influenced by the mean radio download or uplink rate (in contrast to mobile video QoE [85]). Moreover, higher radio signal strength (RSSI) does not correlate with higher web QoE, suggesting that web QoE is not power limited. Further, we establish which radio-level metrics are strong indicators of QoE, and which should not be relied upon.

[1]We use the term *clicks* to refer to mobile "taps" as well as traditional mouse "clicks."

- We capture the complex relationships between the various network parameters and user experience using *intuitive* and *accurate* machine-learning models. Given only radio network characteristics, which are available to network operators even without traffic monitoring, our model can predict the web QoE with accuracy as high as 84%, improving accuracy by 20% compared to the obvious baseline. Network operators can use this model to continuously monitor and improve web QoE by adjusting network parameters.

The rest of the chapter is organized as follows. In Section 4.1, we present the background and details of our data collection process. In Section 4.2, we discuss related work. In Section 4.3 we describe and evaluate our approach for estimating user experience metrics from network traces. In Section 4.4 we present a characterization of how different network parameters affect web browsing experience. We develop a unified web QoE model for web browsing experience and present our findings in Section 4.5. We conclude in Section 4.7.

4.1 Background

Mobile network users care about the web experience rather than individual network metrics such as throughput and latency. Thus, cellular carriers have a significant interest in using their infrastructure to measure and improve web QoE rather than traditional network metrics, especially when there are trade-offs. To better understand the challenges in measuring web QoE, this section first provides a brief overview of the cellular network architecture, focusing on most relevant aspects for our study, the datasets we use, and the applications of web QoE.

4.1.1 Cellular Network Architecture

A Universal Mobile Telecommunication System (UMTS) is a 3rd Generation (3G) mobile data network, consisting of two major components: a Radio Access Network (RAN) and a Core Network (CN). The RAN includes user equipment (UE), base transceiver stations (i.e., NodeBs), and Radio Network Controllers (RNCs). The CN consists of Serving GPRS Support Nodes (SGSNs) and Gateway GPRS Support Nodes (GGSNs). A UE is a mobile device (smartphone, 3G card, etc.) that connects to the NodeB over the radio channel.

Each base station has multiple antennas (typically 3-6), each of which provides radio coverage for an area called a cell sector, which has a particular frequency and other channel characteristics. The primary cell sector is periodically selected based on the signal strength information, while the UE maintains connections to a set of sectors in range called the active set. The traffic to and from the UE is sent to the corresponding NodeB, via RNC, which controls multiple NodeBs, schedules transmissions, and performs all Radio Resource Control (RRC), such as signaling, handovers, and assignment of Radio Access Bearers (RABs).

Within the CN, an SGSN transfers data between RNC and GGSN on behalf of the UE. A GGSN acts as a packet gateway and router between the cellular network and external networks such as the Internet. A GGSN also maintains IP connectivity between UEs and external IP networks.

4.1.2 Data Collection Apparatus

Mobile operators often collect metrics derived from the traffic that passes through network elements in order to manage the network. For example, radio statistics such as Received Signal Strength Indication and handovers are often collected from RNCs and end-to-end throughput and latency metrics are often derived from measurements in the GGSN. This study is interested in whether such low-level network measurements can be used to measure and understand mobile web QoE.

Thus, for the purposes of this study, we simultaneously collect two anonymized data sets, HTTP transaction records from the interfaces between GGSNs and SGSNs, and radio data from a set of RNCs. The datasets cover a major metropolitan area in the western United States over the duration of one month in 2012.

The HTTP records contain IP flow-level information for web-browsing sessions, and it includes items like client and server IP addresses and TCP ports, flow duration, anonymized device identifier (IMEI), bytes transferred, and TCP flags. Also included are relevant HTTP headers, which include information on URL, user agent, content type, content length etc. The query parameters in URLs are anonymized via hashing. The radio data contains event-level information for each anonymized user. For example, this data includes RRC measurement reports that periodically report the RSSI, signal to noise ratio, etc. of each UE to the RNC, handover events, RRC throughput utilization, etc. The signal strength and throughput measurements are reported every 2 seconds. Other measurements are reported based on discrete event level data (e.g., when a handover happens, when user connects, disconnects etc.). A full list of events that we use is in Section 4.4.

Throughout this study, our analysis focuses on three leading mobile websites (*News*, *Social*, *Wiki*) that consistently appear in the top 100 websites [31]. We also confirm our observations by doing preliminary analysis on five other top websites. Our one month long HTTP trace contains information on 2 million web sessions to these 3 websites comprising 70 million HTTP requests and around 1 million different UEs. Our radio dataset contains complete information about 100,000 of these sessions.

We emphasize that all the device and user identifiers are anonymized before any analysis is conducted in order to protect privacy. In addition, the outputs of models in this study are aggregated (e.g., per region and/or network element), so it does not permit the reversal of anonymization or re-identification of users.

4.1.3 Applications of Web QoE Model

Mobile operators monitor network metrics for several purposes, and the ability to monitor web QoE would complement the same applications. First, continuous measurement of metrics permits early detection of network problems. Monitoring web QoE would permit operators to prioritize problems that have the most impact on actual user experience and the understanding of how network factors influence web QoE would help troubleshooting. Second, trending network metrics is invaluable for capacity planning, as they provide objective benchmarks to measure the relationship between investment in infrastructure, such as base stations, and user experience. Third, cellular networks are extremely complex to manage and optimize, involving a huge

amount of parameter tuning in the RAN and CN. The adjustment of these parameters often involve implicit trade-offs between different aspects of network performance, such as average capacity vs. peak latency. Monitoring web QoE and understanding how various network factors influence it provide an objective way for operators to perform such parameter optimizations.

4.2 Related Work

Web traffic modeling: The most widely used technique to model web traffic and identify web pages from network traces is based on idle time [97]. It has been used extensively for characterizing web traffic in several works [60, 109]. This approach works well for simple static web pages. However, it does not work well for most web pages today since they include dynamic content (shown in Section 4.3). To overcome this limitation, a page detection algorithm that works for dynamic content was proposed [84]. However, it only identifies clicks resulting in new web pages and does not identify clicks within a page. We propose and evaluate a text classification-based mechanism that has high accuracy in identifying user clicks in Section 4.3.

Web Performance Studies: There have been several efforts made in previous works towards improving web performance. These include developing better browsers specifically for mobile devices [100], techniques to optimize webpages [14, 16], and detecting inefficiencies in HTTP [59, 102]. More recent work has characterized how web site complexity can affect user experience [62]. Unlike these past works, the focus of our work is on understanding the impact of cellular radio characteristics on mobile web browsing sessions with the aim of helping network operators make informed choices on improving web QoE.

Performance of network protocols over cellular networks: Past work has also looked at the performance of TCP and HTTP on LTE network highlighting the need to develop more LTE-friendly transport and application protocols [47], characterized the physical and MAC layers in CDMA and its impact on TCP performance [45], studied how large buffers in cellular networks cause TCP queing delays [46]. These efforts have helped understand and improve transport layer and application performance over cellular network, and hence user experience indirectly. In this work understanding the impact of transport layer protocols on user experience is not our immediate focus—the goal of this study is on understanding how radio network parameters impact user experience.

Measures of web browsing user experience: User experience studies in the past have shown that a complete page load time has an impact on user satisfaction [61, 69, 77]. These works are primarily based on controlled studies with few users, and they involve logging page load times and user feedback using client-side instrumentation techniques. However, since network operators do not have access to client-side logs, it is challenging to exactly measure the page load time. However, in our traces we observe that large fraction of the pages are only partially downloaded and we define the partial download ratio metric to capture user experience. Similarly, past work has also extensively used several metrics related to user browsing behavior to quantify user satisfaction including user clicks, dwell time and scrolling [80, 93]. We also use metrics related to user click behavior. However, since we do not have client-side instrumentation, we are unable to capture other behavior such as dwell time and scrolling and incorporate them in our study.

Moreover, our work takes a step forward by analyzing the impact of radio network factors on these different user experience metrics.

QoE in other domains: Several past efforts study the impact of network factors on user experience and user satisfaction in other applications. Measured impact of bitrate, jitter, and delay on VoIP call duration is used with a machine-learning approach to derive a user satisfaction metric [49]. Past works have employed machine-learning algorithms to develop predictive models for Internet video user engagement [58, 85]. Radio network factors, such as signal strength and handovers, are used to quantify their impact on video viewing experience [85]. Our work focuses on performing a similar analysis for mobile web browsing experience.

4.3 Extracting User Experience Metrics

Network operators cannot access detailed server-side or client-side logs of user browsing patterns. Hence they need to reconstruct web browsing sessions and estimate user experience metrics from network traces. However, over the years, webpages have evolved from serving relatively simple static objects such as hypertext to serving dynamic and even personalized content. This makes it even more challenging to reconstruct mobile web sessions and extract user activities from network traces alone.

Previous work identified *engagement* as a key measure of user experience because more satisfied users tend to stay around longer and use an application more [49, 58, 88, 112]. For web browsing, two central engagement metrics recognized in the web analytics industry are *session length* (i.e., the number of pages a user clicks through) and *abandonment* or bounce rate (i.e., if a user leaves the website after only visiting the landing page) [67]. Unfortunately, both of these metrics necessitate the identification of user *clicks*, which is non-trivial from the perspective of an operator. The difficulty comes from lack of access to client-side or server-slide logs and HTTP records do not readily distinguish requests that are initiated automatically by the browser (e.g., embedded objects) and those that are initiated by user activity.

In Section 4.3.1, we present and evaluate a novel click detection algorithm based on machine learning, which achieves higher accuracy than the best known approach. Then, in Section 4.3.1, we extract different user experience metrics from the dataset using our algorithm. Based on our findings, we propose *partial download ratio* as a more fine-grain metric that more precisely captures user experience impairments due to network conditions, as opposed to user interest.

4.3.1 Detecting Clicks

Limitations of Previous Techniques

The most common approach for differentiating between clicks and embedded objects using network traces is to use the idle time between requests [60, 97, 109]. This approach is based on the assumption that the idle time for requests for embedded objects will be extremely short since they are automatically generated by the browser, whereas requests generated by clicks would typically have higher idle time since they require manual intervention. Therefore, these techniques use a pre-defined threshold and classify a request as embedded object if and only if the idle time

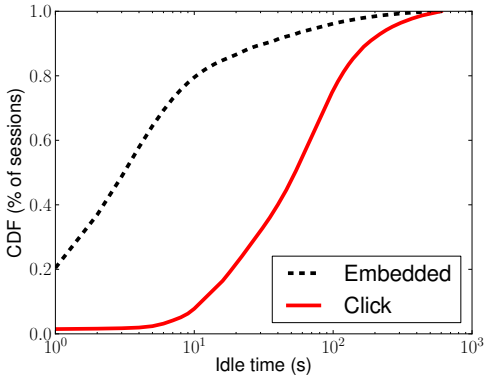


Figure 4.1: CDF of arrival time for clicks vs embedded objects

is shorter than the threshold. However, we find that in modern mobile web pages, a non-trival fraction of embedded objects have idle times as long as many user clicks (e.g., requests generated by periodic beacons from third-party analytic services [84]). For example, Figure 4.1 shows the distribution of arrival times for next click and next embedded objects. This figure is based on web requests from around 50,000 web sessions on the three websites. We labeled each of the web requests in these sessions as clicks or embedded objects manually. An idle time threshold approach would select a point on the x-axis and classify all objects to the left as embedded and those to the right as clicks. We see that there is no idle time threshold that we can select that achieves lower than 20% error on at least one of the two classes.

To improve on this approach, StreamStructure [84] exploits the structure of "desktop" web pages to detect requests for new webpages. However, we show in the next section that it is not as adept at identifying clicks in mobile web pages. Moreover, it is a page detection algorithm that is used to identify clicks resulting in new pages. Other client-side interaction (e.g., clicking to play a video within a page) are not identified by this algorithm.

Our Approach

Our approach to differentiate between clicks and embedded objects is based on our observation that most of the embedded objects are hosted by third party services such as advertising agencies, Content Distribution Networks (CDNs) and analytics services [62]. This opens up an opportunity to distinguish embedded objects from clicks by inspecting request URLs. For example, a request to googleanalytics.com is *very likely* to be an embedded object, while a request to news.google.com is *very likely* to be a click. Hence we can employ text based classification algorithms that have been extensively used in other domains (such as spam filtering [107] and sentiment analysis [78]) to classify requests. We would need to learn the classification model separately for each website/domain. In the remainder of the section, we explain four steps in our approach.

Step 1: Grouping Sessions: We first filter our sessions to a specific website. We only study traffic originating from web browsers, and hence filter out traffic of native mobile apps using User Agent HTTP header. Similar to the approach in StreamStructure [84], we group requests

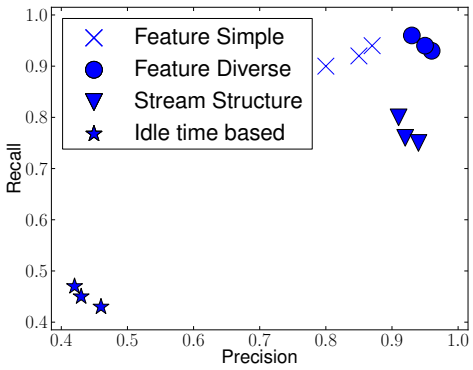


Figure 4.2: Our approaches have higher precision and recall compared to previous approaches

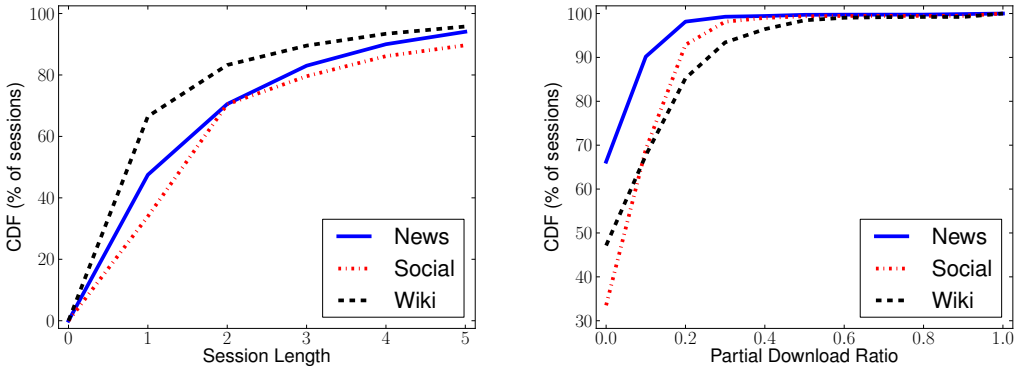
into different sessions using the anonymized IMEI and Referer header information. The IMEI information helps us separate sessions from different devices. The Referer field identifies the address of an object from which the new request came from. The first request in a session has an empty Referer field. The Referer field is further used to build the request chain within a session. It also helps us separate simultaneous requests from multiple browser instances from the same user equipment.

Step 2: Extracting Features: In order to perform text classification, we extract features from the requested URLs. We extract two sets of features to represent the URLs:

- Feature simple: We extract a bag of words [101] from the domain name. For example, the feature set for the URL www.blog.xyz.com/my/blog/abc.html is <blog, xyz, com>.
- Feature diverse: In addition to domain name features, we include features from the URN and type of content. Hence, the feature set for the above URL would be domain = <blog, xyz, com>, urn = <my, blog, abc.html> and type = html.

Step 3: Automatic Labeling to Obtain Training Set: In order to create a classification model to label the URLs as clicks and embedded objects, we need a training set labeled with the ground truth. To obtain the training set, we inspect only the very first 10 seconds of every web session. We assume that only the first request during this time frame was a click and the remaining requests are for embedded objects and collect both feature simple and feature diverse along with the ground truth based on the assumption. We pick 10 seconds because based on the ground truth in Figure 4.1, almost all user clicks have an idle time of more than 10 seconds and almost 80% of the embedded objects are requested with this time frame. This automatic labeling technique enables our approach to be applied to any website without any manual labeling or ground truth.

Step 4: Running Classification Algorithm: We first learn the classifier model using the training set, and then input the entire dataset and classify each request as the click or embedded object. After testing with multiple machine learning algorithms (such as decision trees, logistic regression, Support Vector Machines [101]), we found that Naive Bayes performs the best compared to other approaches. This is not surprising given that Naive Bayes has been found to perform the best in other text classification problems as well [38].



(a) Session Length (b) Partial Download Ratio

Figure 4.3: CDF of user experience metrics

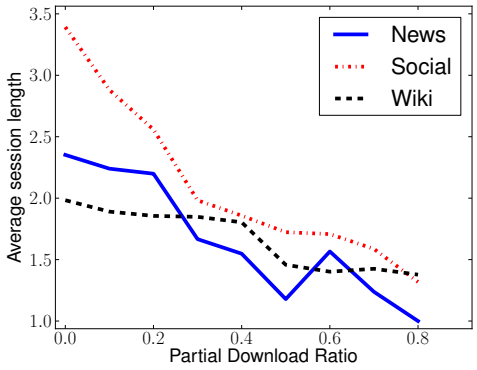
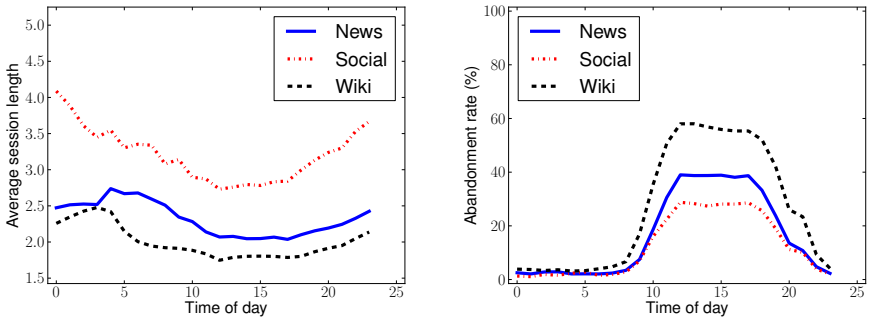


Figure 4.4: Session length decreases with increasing partial download ratio

Validation

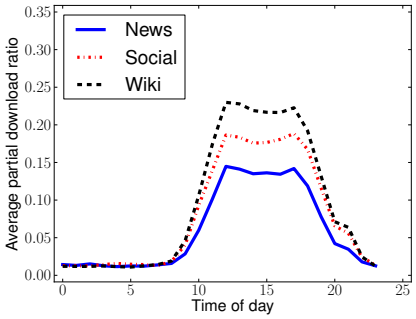
To validate our approach, we apply it to the three web sites we study. For each web site, we manually inspect its structure in detail in order to label each HTTP request as either a click or a as an embedded object. We manually label one day of trace data with roughly 50,000 sessions.

We then compare the idle-time threshold based approach, StreamStructure and our approach (both using Feature simple and Feature diverse) and then estimate the performance in terms of both precision and recall. Precision is defined as the number of correct clicks identified by the total number of clicks identified, and recall is defined as the number of correct clicks identified by the total number of clicks. Figure 4.2 shows the precision and recall using each the three websites. Our Feature simple and Feature diverse have higher recall than the previous baselines. Feature diverse has higher precision than Feature simple because some embedded objects are hosted on the main domain. Feature simple will incorrectly classify this object as a click since it just uses features from the domain name. In the remainder of this chapter, we hence use the Feature diverse approach to identify clicks.



(a) Session Length

(b) Abandonment Rate



(c) Partial Download Ratio

Figure 4.5: Time of day effects on the experience metrics

4.3.2 Measuring User Experience

As described earlier, session length and abandonment rate are two important metrics that the web industry recognizes as representative of user engagement. Our click detection algorithm presented above enables operators to estimate session length and abandonment rate using only HTTP traces collected from the network. However, session length and abandonment rate are relatively coarse engagement metrics because they are also influenced by user interest, which is a confounding factor that is difficult to measure for both network and website operators. Indeed, we find that many web sessions are only one click (and thus, by definition, abandoned). These metrics do little to distinguish satisfied and dissatisfied users of these single-click sessions. In this section, we show that *partial download ratio*, i.e., the fraction of HTTP objects that are not completely downloaded in a session, is strongly correlated with session length and abandonment rate, so we can use it as a proxy to estimate user experience, even for sessions lasting a single click.

To better understand the session length, abandonment rate, and partial download ratio metrics, we extract web sessions and estimate number of clicks for each session for the three different websites from the entire 1 month HTTP record trace. Figure 4.3(a) shows the distribution of session lengths. We observe that all sessions have length less than 10 on all the three websites. A significant fraction the sessions on all the three websites have a length of 1 (47% for News,

33% for Social, 67% for Wiki). The overall abandonment rate is 35% for the three websites. These observations highlight the need for a user engagement metric that can highlight network problems in sessions of length one.

One such candidate measure in common use is the web page load time (i.e., the time it takes to load a page). However, it is known that web page load time is difficult to measure from HTTP traces because the traces do not capture the browser's rendering pipeline [111]. Moreover, without a priori knowledge of web page structure, operators can not easily distinguish complete vs. incomplete page loads. Therefore, naïvely using download time to approximate page load time would incorrectly suggest that abandoned pages have low load time.

Instead, we propose that partial download ratio is a useful proxy metric for user engagement. Figure 4.4 shows the average session length as a function of the partial download ratio. We see that there is roughly a negative linear relationship between the partial download ratio and session length, supporting our hypothesis that users are less engaged when more objects on the page fail to load completely (or do not load completely before the user moves on). The linear coefficient is different for each website, as website design likely influences how much partially downloaded content effects the user experience, but the coefficient can be easily learned using regression. For example, using a linear fit to determine session length in terms of partial download ratio, the partial download ratio co-effieceints for the News, Social, and Wiki websites are -1.6, -2.36, -0.85 respectively. Figure 4.3(b) shows the distribution of partial download ratios for each session. We also observe that over 60% of the sessions have objects that are partially downloaded on each website.

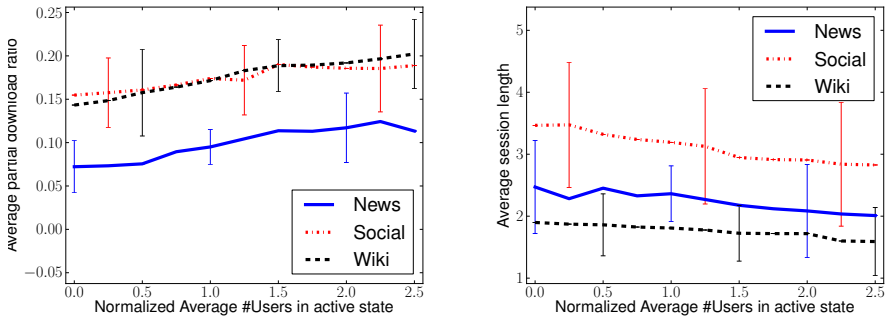
Figure 4.5 shows the average session length, abandonment rate, and partial download ratio by time of day. We observe strong temporal patterns in the user engagement metrics. Lower session lengths, higher abandonment and higher partial download ratio occur during peak hours (10 am - 6pm) compared to the rest of the day.

Corresponding to the different linear coefficients we see in Figure 4.4, we observe that the web QoE metrics are different across different websites. This is likely because, as previous work has showed [62], user experience is dependent on factors other than network quality, such as how mobile-friendly the website is, the number of objects, type of objects etc.

4.4 Analyzing Network Factors

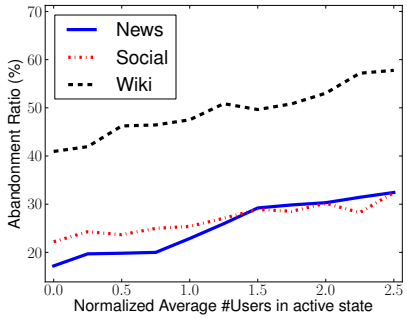
Our first goal is to understand the relationships between individual network factors and web QoE, with the end goal of building models that can be used to improve web QoE by tuning various network parameters. We first itemize all radio network factors that we study, which may effect web QoE:

- *Number of soft handovers (SOHO):* A soft handover occurs when a cell sector is added or removed from the set of cell sectors that a UE is connected to. A SOHO is a "make-before-break" handover in that the link to the new sector is established before an old link is removed. From radio network data that we collected from the RNCs, we count the total number of soft handovers during a session.
- *Number of inter-frequency handovers (IFHO):* An IFHO occurs when the UE switches to a cell sector that is operating on a different frequency. An IFHO is a "break-before-make"



(a) Partial Download Ratio

(b) Session Length

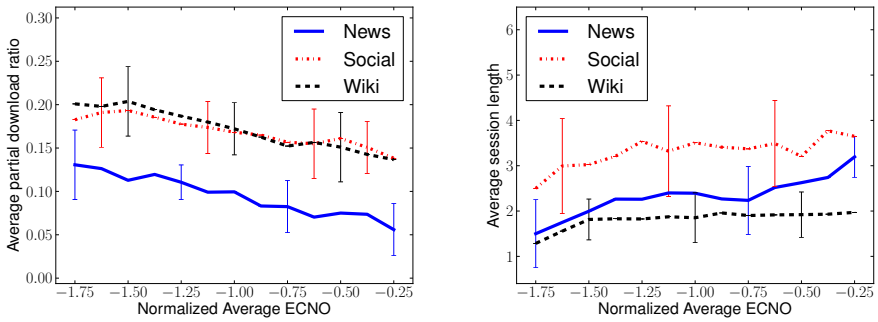


(c) Abandonment Rate

Figure 4.6: Higher load in the cell (measured in terms of number of active users) leads to worse web QoE. Session length has higher variance since it is a more "noisier" metric as explained in Section 4.3.2

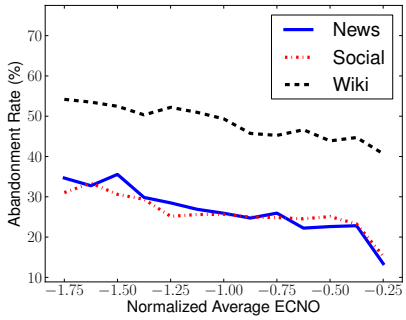
handover because a device can only listen on one frequency at a time; thus, it must break the link with all old sectors before establishing the link with the new one. We count the number of inter-frequency handovers during a web session.

- *Number of inter-radio access technology (IRAT) handovers:* An IRAT handover happens when a UE switches between different radio access technologies (e.g., UMTS to GPRS or EDGE). These do not include handovers to and from WiFi since our data collection apparatus does not capture such handovers. An IRAT handover is also a "break-before-make" handover because the device must disconnect entirely from the current radio network before connecting to the new one. This process involves a significant amount of network signaling and can take several seconds.
- *Number of admission control failures (ACF):* We count the number of times the UE fails to complete the admission control procedure during the web browsing session. These events mostly occur when the radio network is overloaded.
- *Number of RRC failures (RRC):* An RRC failure occurs if the RNC is overloaded and it cannot allocate a request from the UE for more radio resources. We count the number of RRC failures within a web session.



(a) Partial Download Ratio

(b) Session Length



(c) Abandonment Rate

Figure 4.7: Higher signal energy to interference (ECNO) leads to better web QoE

- *Average Received Signal Code Power (RSCP):* This is the downlink power received by the UE receiver on the pilot channel. It is measured in dBm.
- *Average received energy per chip of the pilot channel over the noise power density (ECNO):* It is expressed in dB and it measures how well a signal can be distinguished from the noise in a cell. It is measured in dB. Note that ECNO is measured on the pilot channel and thus may be different from the SINR of the traffic channel.
- *Average received Signal Strength Indicator (RSSI):* Expressed in dBm, it is the wide-band received power within the relevant channel bandwidth. It is related to RSCP and ECNO as follows: RSSI = RSCP - ECNO. Note that RSSI is measured on the pilot channel and thus may be different from the received power of the signal on the traffic channel.
- *Average uplink and downlink radio data throughput:* We compute the average uplink and downlink data rates for the UE when it is in active state during the web session in Kbps. Note that the radio data rate is not equivalent to the long-term throughput because it is only measured when the device is scheduled to send/receive data (the radio link is time and code division multiplexed). The radio data rate does not count the non-scheduled time slots in the denominator of the throughput calculation. The number of users in active state (see below) serves as an estimate of the number of competing flows, as a sector schedules each user in a proportionally fair manner.

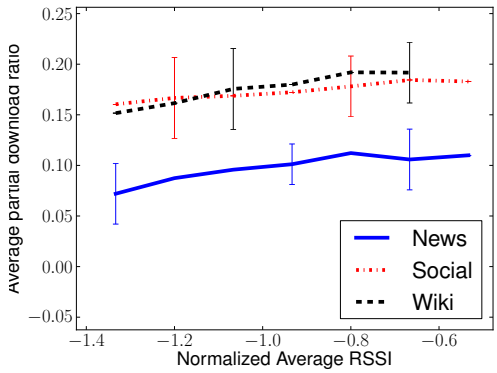


Figure 4.8: Surprisingly, higher received signal strength leads to higher partial download ratio

- *Average number of users in active state:* We measure the number of active users served by each cell at a minute-level granularity. Using this information we compute the average number of users that are served by the cell that the UE is connected to during the web session. This is an indication of the load in the cell.

We report the normalized the value of RSCP, ECNO, RSSI, average uplink and downlink throughput and number of users in active state as a fraction of the mean of the metric. For example, instead of reporting the absolute value of RSSI, we report (RSSI - mean(RSSI))/mean(RSSI) and hence the plots can be read as x% above or below the average.

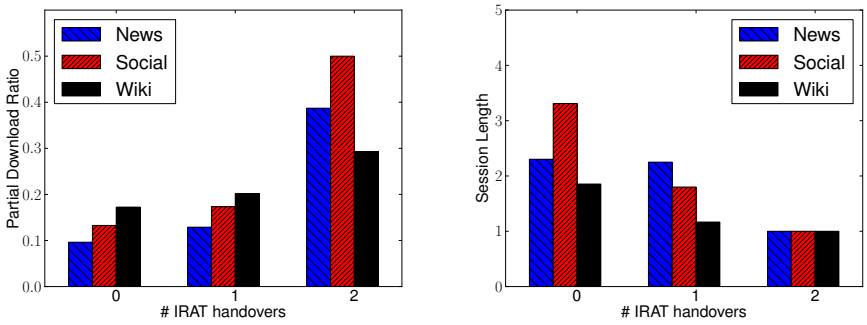
4.4.1 How network factors impact web QoE

To understand how each network factor individually affects web QoE metrics, we plot web QoE metrics against measured values of network factors from our radio data. The main takeaways are as follows:

1. Higher network load results in worse web QoE. Number of users in active state in a cell is an indication of load in the network. As Figure 4.6 shows, there is a linear relationship between the load and various web QoE metrics. For instance, adding 25% more users than the average can increase abandonment by 2 full percentage points. Increasing cell load also leads to lower session lengths and higher number of partially downloaded objects on average. This relationship between load and web QoE metrics holds even when conditioned by time-of-day, though cell load is significantly higher during peak hours (not shown due to space constraints). The results suggest that cellular network operators can improve web QoE by decreasing cell load by deploying more radio cells or re-distributing users across cells.

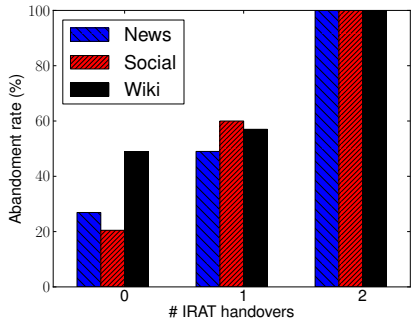
2. Higher signal strength (RSSI) does not necessarily correlate with better user experience, but higher signal energy to interference (ECNO) does. As Figure 4.7 shows, increasing the signal energy to interference (ECNO) by 10% above average reduces abandonment rate by about 2 percentage points, increases average session length between 2.6% and 9.4% and improves partial download ratio by 0.7 percentage points. In contrast, Figure 4.8 shows that sessions with higher RSSI have higher partial download ratio on average.

These results confirm that, similar to recent WiFi findings, ECNO (an analogous measure



(a) Partial Download Ratio

(b) Session Length



(c) Abandonment Rate

Figure 4.9: IRAT handovers have a strong impact on web QoE—all sessions with 2 handovers are abandoned.

to the SINR of WiFi beacons) is a better indicator of channel quality than RSSI because RSSI does not exclude the power of noise and interference. This finding suggests that web QoE is interference and noise limited, not power (i.e., coverage) limited. We did not observe any impact of RSCP on user experience metrics (not shown).

3. IRAT handovers lead to worse web QoE. IRAT handovers had the strongest impact on user experience, as seen in Figure 4.9. Sessions with IRAT handovers are much shorter than those without IRAT handovers. Also, all sessions with more than 1 IRAT handover were abandoned. The impact of other handovers (soft handovers, inter-frequency handovers) and failure events (access control failure, RRC failure) on web QoE were negligible. Figure 4.10 shows that increasing number of such handovers and failures leads to minimal increase in partial download ratio. This indicates high robustness against these types of events, hence they should not be used to assess web QoE and their management would likely yield insignificant improvement.

4. Higher radio data rate does not necessarily lead to better web QoE. Figure 4.11 shows the impact of radio data rate on partial download ratio. As web objects are primarily downloaded onto the mobile device, we start by looking at the downlink direction and find that higher data rates do not improve partial download ratio (Figure 4.11a). As expected, uplink data rate shows no impact (Figure 4.11b). We find similar relationship between data link rates and other web QoE metrics (not shown). While it may not be intuitive that data rate and web QoE metrics

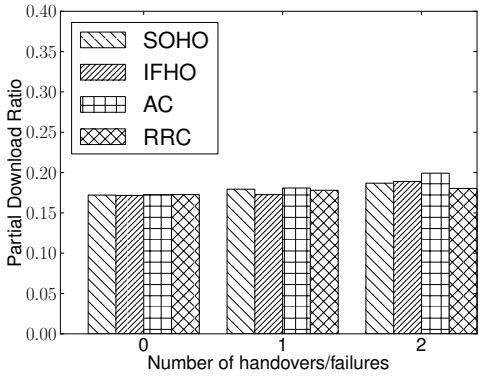
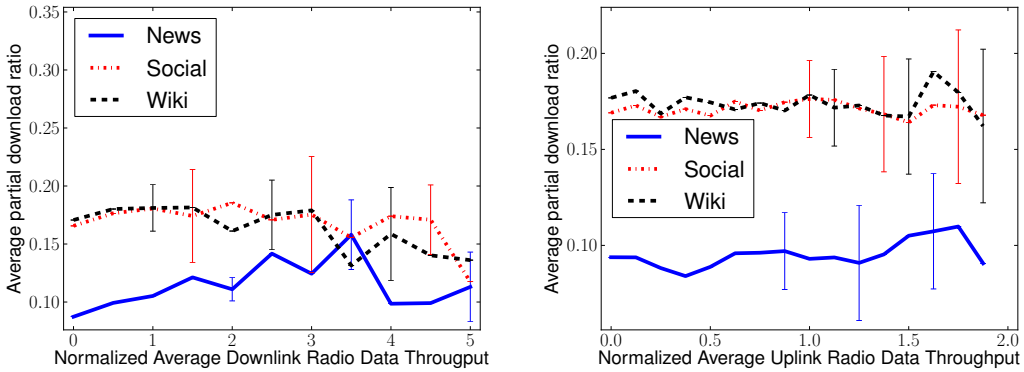


Figure 4.10: The impact of soft handovers, inter-frequency handovers, access control failures and RRC failures on web QoE is minimal



(a) Downlink (b) Uplink

Figure 4.11: Radio data link rate does not impact partial download ratio

have weak relationship, it has been shown that web browsing traffic is more latency-limited than throughput-limited [21, 44].

4.4.2 Analysis on Other Websites

To test whether the observations we made above hold for other websites, we analyze one day's worth of HTTP records and radio data for five other leading mobile websites (*Shopping*, *Marketplace*, *News2*, *Social News* and *Blog*) that consistently appear in the top 100 [31].

In Table 4.2, we characterize these websites based on two key metrics of website complexity that past work has identified [62]— namely, (1) the average number of objects requested per click and (2) the average number of domains from which requests are served. We found that these websites represent a varied range both for complexity and for user behavior. For example, *News*, *News2*, *Social* and *Blog* have the highest complexity these metrics, whereas *Shopping* and *Marketplace* are less complex. Moreover, users tend to have different browsing behavior on these websites: *Shopping*, *Marketplace*, and *Social News* sites understandably tend to have higher session lengths, while *Wiki* and *Blog* tend to have low session lengths.

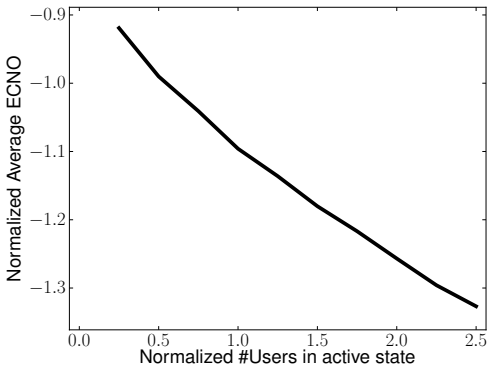
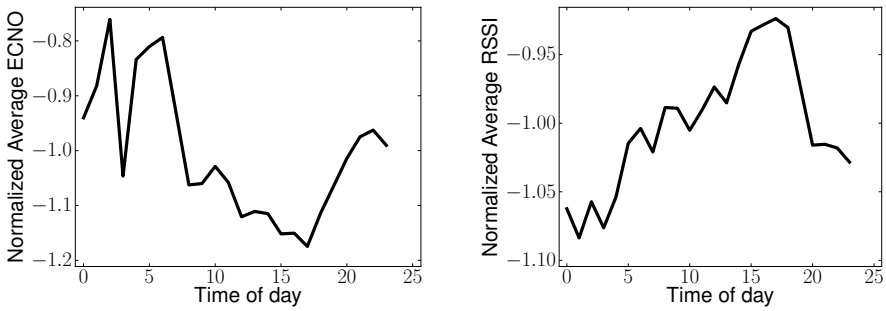


Figure 4.12: Number of users vs ECNO



(a) ECNO (b) RSSI

Figure 4.13: Time of day effect on signal strength parameters

Table 4.2 shows that our result hold for this varied set by correlating the impact of increasing each of the radio network factors on partial download ratio. For each radio network factor (e.g., RSSI, ECNO, etc.), we tabulate the slope of the partial download ratio vs. the network factor. For each of RSSI, ECNO, # IRAT handovers and # Users, the partial download ratio graphs exhibit the same trend on increasing the radio network factor even across the varied set of websites. For example, increasing ECNO decreases partial download ratio (i.e., negative slope across all sites).

# IRAT handovers	Normalized ECNO	Normalized RSSI
0	-1.09	-0.97
1	-1.53	-1.15
2	-1.84	-1.21

Table 4.1: We observed lower average ECNO and RSSI for sessions with IRAT handovers

	Website complexity		Average webQoE		
Website	# Domains	# Objects	Average P.D.R	Average Session Length	Abandon Rate (%)
News	13.4	23.1	0.1	2.2	21
Social	13.1	20.2	0.15	3.1	23
Wiki	3.69	13.31	0.16	1.8	41
Shopping	4.6	7.5	0.12	8.5	10
Marketplace	3.2	3.6	0.04	12.3	5
News2	15.7	29.9	0.09	3.4	15
Social News	11.65	7.38	0.03	10.9	8
Blog	11.89	20.40	0.17	2.65	30

	Impact of increasing radio factor on Partial Download Ratio (P.D.R.)			
Website	RSSI	ECNO	# IRAT	# Users
News	0.0006	-0.007	0.15	0.0004
Social	0.0004	-0.004	0.06	0.0003
Wiki	0.0011	-0.005	0.18	0.0003
Shopping	0.0015	-0.004	0.07	0.0003
Marketplace	0.0007	-0.001	0.06	0.0002
News2	0.0010	-0.008	0.17	0.0004
Social News	0.002	-0.005	0.12	0.0003
Blog	0.0003	-0.009	0.19	0.0005

Table 4.2: Observations made in Section 4.4.1 hold for a varied set of websites

4.4.3 Comparison with Other Mobile Applications

Our findings that higher load (in number of users) and lower signal to noise ratio (ECNO) correlate with lower web QoE is not entirely surprising and confirms previous findings on the relationship between these network factors and QoE of mobile video streaming [85]. Interestingly, as in the case of video streaming, the relationship between these two network factors and abandonment rate are both linear and have roughly the same slope.

In contrast to findings on video streaming, however, we observed that only IRAT handovers were disruptive to web QoE and that web QoE metrics were uncorrelated with SOHOs and IFHOs. This finding suggests that web browsing is more tolerant to minor handover disruptions than video streaming. IRAT handovers are much more disruptive because changing radio technologies can take several seconds to complete, which is long enough to influence user perceived latency. Moreover, we find that, unlike mobile video streaming, the radio data rate is uncorrelated with web QoE metrics. This may be because video streaming is a more bandwidth intensive application, whereas web browsing is more sensitive to latency.

In summary, our findings complement previous work on cellular mobile video [85], demonstrating that reducing load and improving ECNO are equally important for both applications. However, carriers need not optimize handovers (except IRAT) or radio throughput rates if they only want to improve web QoE.

4.4.4 Dependencies and Other Factors

We found that many network factors under study are not independent of each other. An obvious example is that RSSI is related to ECNO and RSCP (as we mentioned earlier). We also found several other dependencies between the radio network factors. Some examples are:

- The number of users in active state in a cell and ECNO are dependent on each other [30]. As shown in Figure 4.12, there is a linear relationship between the two—adding more users into the cell steadily decreases ECNO.
- Table 4.1 shows that sessions that experience IRAT handovers also experience lower signal strength (RSSI) and lower signal energy to interference (ECNO).

Further analyzing the radio network factors, we also observed significant time of day effects. Figure 4.13 shows the average value of RSSI and ECNO observed per hour of the day over the entire one month dataset. We observe that average signal strength to interference (ECNO) is lower during peak hours compared to non-peak hours. On the other hand, average signal strength (RSSI) is higher during peak hours compared to non-peak hours. In Section 4.3, we also observed strong temporal effects on the various user experience metrics (Figure 4.5). These could also be caused by external factors/reasons—for example, users are less likely to engage in long browsing sessions during working hours pointing to the need for including external factors such as time of day into the analysis.

In summary, complex interdependencies between network factors as well as external factors (e.g. time of day) make it very challenging to understand and quantify the true impact of each network factor using correlation analysis. This points to the need to use more systematic techniques, including machine learning algorithms, to capture the complex relationships in order to quantify the impact of network factors.

Model	Avg. Accuracy (%)
Radio factors alone	73.02
Radio factors + time of day	79.25
Radio factors + time of day + website	83.95

Table 4.3: Adding time of day and learning a separate decision tree for each website improves accuracy.

4.5 Modeling Web QoE

Our end goal is to develop models that can be used by cellular network operators to improve web QoE by tuning network parameters. For example, our model should help answer the question "how much can we improve the partial download ratio if we increase ECNO by 1 dB?". The model should also help network operators monitor web QoE metrics using only radio network characteristics. To achieve this goal, the QoE models that we build should be *intuitive*, *accurate*, and must be able to *predict web QoE from network factors alone*.

Building an accurate QoE model is challenging because of the complex relationships between network factors and web QoE metrics, interdependencies between various network factors, and also due to external factors (e.g., differences between websites, time of day effects). To tackle these challenges, we use machine learning to capture the dependencies and relationships, and develop and evaluate models that can predict web QoE metrics. The model we derive will express web QoE metrics as a function of radio parameters; specifically, we wish to capture the relationship:

$$WebQoE = f(RadioNetworkParameter_{1..n})$$

where $WebQoE$ denotes one of the user experience metrics (partial download ratio, abandonment, or session length), and
$RadioNetworkParameter_i$ denotes the ith observed radio network parameter listed in Section 4.4.

4.5.1 Evaluation

We predict: (1) partial download ratio, (2) session length, (3) if the session includes partially downloaded pages or not (*part* or *full*), and (4) if the user will abandon a session or not (*abandoned* or *not-abandoned*). The choice of the machine learning algorithm is important because the model it learns should be expressive enough to capture all the complex relationships and dependencies. After experimenting with different regression, tree and bayes algorithms (such as linear and logistic regression, variations of decision trees and naive Bayes) we found that linear regression worked best for (1) and (2), and C4.5 decision tree algorithm was able to predict the binary classification most accurately for tasks (3) and (4). We use 10-fold cross-validation to evaluate all our models [101].

Evaluating Linear Regression Models: Since most network factors have a linear relationship with web QoE metrics (session length and partial download ratio) and since they are linearly

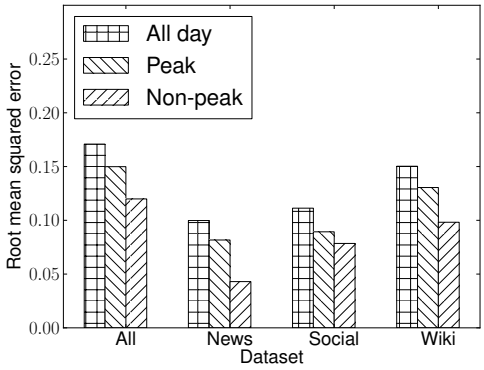
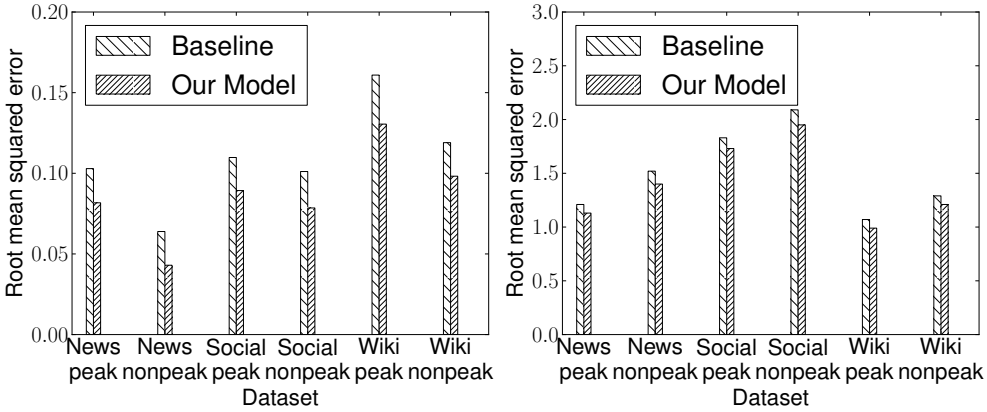


Figure 4.14: Learning a separate regression models for each website and time of day (peak/non-peak) improves accuracy.



(a) Partial Download Ratio

(b) Session Length

Figure 4.15: Our models are more accurate than the baseline in predicting partial download ratio.

dependent on each other (Section 4.3), linear regression is well-suited for capturing these relationships. To measure the "goodness-of-fit" of linear regression, we use the standard measure of root mean squared error (RMSE), where lower RMSE values indicate better prediction. Figure 4.14 shows RMSE when using linear regression to predict partial download ratio. We show the results for each website separately on the x-axis, as well as the results for data from all web sites taken together ("*All*"). The three bars in each cluster represent RMSE for the entire day, for peak hours only, and for non-peak hours only—each of these have separately learned models due to significant time of day effects. We see that separate models for time of day and individual web sites results in significant prediction improvement, as indicated by lower RMSE values.

We repeated these experiments for session length and found similar improvements from splitting data and learning models for each split (not shown). Hence, our final linear regression models for both partial download ratio and session length are *each* a collection of six linear regression models: one each for each combination of web site (*News*, *Social*, *Wiki*) and time of day (*Peak*, *Non-peak*).

	Partial		Abandonment	
Dataset	Model	Baseline	Model	Baseline
News	80.4	59.7	78.6	75.6
Social	87.7	72.5	82.0	78.6
Wiki	80.6	70.3	62.3	53.3

Table 4.4: Our models are more accurate than the baseline in predicting partial downloads and abandonment.

Dataset	# active users	RSSI (dBm)	ECNO (dB)	# SOHO	# IRAT	Constant
News Nonpeak	0.0002	0.0005	-0.0043	0	0	0.0411
News Peak	0.0002	0	-0.0032	0	0	0.0976
Social Nonpeak	0.0002	0.0005	-0.0037	-0.0007	0.0639	0.0485
Social Peak	0.0002	0.0005	-0.0047	-0.0005	0.0627	0.1367
Wiki Nonpeak	0.0002	0.0003	-0.0042	-0.0005	0.0871	0.0848
Wiki Peak	0.0002	0.0004	-0.0037	-0.0004	0.0799	0.2022

Table 4.5: Linear regression coefficients of the model that predicts partial download ratio.

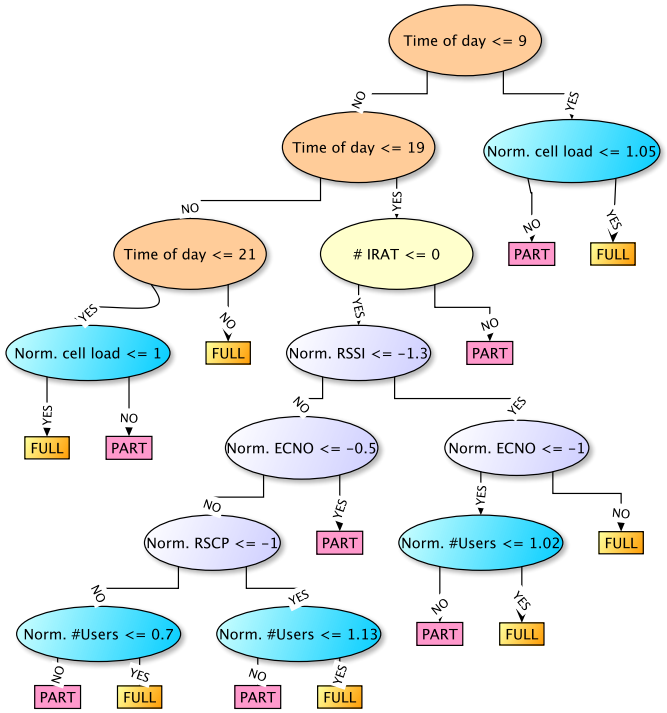


Figure 4.16: Pruned decision tree that predicts partial downloads

We compare the performance of our models with a baseline model that always predicts the mean (using the ZeroR classifier [43]) for the particular dataset in Figure 4.15. In the case of partial download ratio, our model has around 20% lower RMSE compared to the baseline.

For session length, our model has up to 10% lower RMSE. Predicting session length is not as accurate as predicting partial download ratio because session length is more affected by external confounding factors (e.g. user interest), which are very difficult to capture from network traces.

Evaluating Decision Tree Models: We also develop models that make binary predictions for users' web sessions, such as "will this session have partially downloaded pages?" and "will the user abandon this session?". The C4.5 decision tree algorithm performed the best in predicting both these classifications. Table 4.3 shows the accuracy for predicting partial downloads. Refining the model by inputting the time of day (in terms of hour of day (0-23)) along with learning a separate model for each website led to around 11% improvement in accuracy. We observed this for the decision tree that predicts abandonment as well. Essentially, our models for predicting abandonment and partial downloads are both a collection of 3 decision trees, one for each website. They take both the radio factors as well as time of day as input.

Further, we compared our model against a baseline model that predicts the majority class using the ZeroR classifier [43] for each dataset in Table 4.4. Our partial download model achieves up to 20% higher accuracy compared to the baseline while our abandonment model achieves up to 10% more accuracy. Again, smaller improvement for abandonment is due to confounding factors like user interest that we cannot measure, but can significantly influence user abandonment.

4.5.2 Insights and Discussion

Fortunately, both linear regression and decision tree algorithms that gave the highest accuracy also generate very intuitive models. They can hence provide insights to network operators towards tuning network factors to improve web QoE. For example, Table 4.5 shows the regression co-efficients for our partial download ratio model. We observe that the features that the models learnt (number of users, ECNO, RSSI etc.) are the same as those that we found to be impactful in Section 4.4. Moreover, the model also ignores factors such as downlink and uplink throughput that we found to not have an impact on web QoE metrics.

Interestingly, the value of the regression co-efficients are similar across the different datasets. This implies that irrespective of the time of day and the website, tuning a particular network parameter has the same impact on partial download ratio. For example, improving ECNO by 1 dB decreases partial download ratio by roughly 0.004 across all times of day and websites. Network operators can hence use this model to understand the true impact of a parameter. For example, comparing the co-efficients, decreasing IRAT handovers and improving ECNO has the highest impact on improving partial download ratio. We also found similar conclusions from analyzing the regression co-efficients for session length (not shown due to space constraints).

Figure 4.16 shows the pruned decision tree that we learnt for predicting partial download for the *Wiki* website. Again, consistent with our analysis in Section 5, the model picks parameters such as Number of users, ECNO, IRAT etc. to branch on, reconfirming the impact of these factors. Further, the decision tree rules separate the data based on time of day into a similar classification that we made for peak vs. non-peak (e.g., Time of day $<= 9$, Time of day > 9 and Time of day $<= 19$, Time of day > 19). We also observe that the feature splits conform with several of our observations. For example, during non-peak hours the partial downloads are

lower (if Time of day > 21, predict *full*). Also if load is higher partial downloads are higher (if Normalized Num User $<= 1.05$, predict *full* otherwise *part*).

4.6 Discussion

Other websites and native apps: In this chapter, we developed and analyzed QoE models primarily for three different websites. However, preliminary analysis in Section 4.4.2 indicates that the general observations also hold for other websites with different complexities. We also found that (1) the same machine learning algorithms (decision trees and linear regression) performed the best across different websites, (2) the model parameters are similar across websites, and (3) the *ordering* of the importance of network parameters is the same across different websites. Furthermore, our methodology is completely automated and can be easily applied to other websites. In this chapter we focused on traffic from web browsers by filtering based on the User Agent string; this analysis does not include traffic from native mobile apps, and applying our techniques to this type of traffic is an interesting direction for future work.

New technologies: Our study is based on traffic collected from a UMTS network. However, we expect our methodology and results to generally apply to newer technologies such as 4G Long Term Evolution (LTE). Radio network parameters in UMTS have analogous parameters in LTE. For example, the number of users in active state in UMTS is related to the number of users in CONNECTED state in LTE. Similarly, number of RRC failures in LTE is related to the number of RRC failures in UMTS. We observed that the number of IRAT handovers has the highest impact on user experience. This for instance might have a lesser impact when switching between UMTS and LTE since the handovers are expected to complete faster. Similarly since web browsing is more latency-limited than throughput-limited, higher throughput offered by LTE may not have a significant impact on web QoE.

Encrypted traffic: Our click detection algorithm uses URL to classify if a web request is a click or an embedded object. With SSL/TLS traffic, the entire user payload including HTTP URL is encrypted and hence we cannot apply the current methodology. However, a network operator can still identify the domain to which an encrypted web request is sent by correlating the IP address of the web request from DNS server logs—the IP address will likely correspond to a prior DNS request/reply pair. Our click detection technique can be tailored to use this information in the future. Also, encryption would not affect our QoE prediction methodology, since it is based on radio statistics and completely independent of traffic encryption.

Limitations: One of the main constraints faced by a network operator is the lack of client-side instrumentation. This makes it difficult to differentiate between the abandonment caused by the lack of user interest from the ones caused by network issues. For example, a user could potentially have abandoned the session due to lack of interest, and yet the network would have delivered all the data. It is impossible to identify such a situation from network logs alone. Similarly, network operators cannot identify cellular-to-WiFi handovers from cellular network traces alone, and would mistakingly mark such handovers as abandonments. Nonetheless, operators are typically interested in aggregate performance trends and changes that signify network issues or improvements. A few false positives or negatives introduced by these limitations are unlikely

to significantly alter the behavior of aggregate metrics.

4.7 Chapter Summary

In this chapter, we presented a large-scale study that analyzed web QoE, such as session length, abandonment rate and partial download ratio, from a "cellular network operator" point of view. Understanding web QoE from a network operator perspective is challenging due to lack of visibility or instrumentation at clients and servers and a priori knowledge of web site structure. We developed and evaluated text-classification-based mechanisms to extract mobile web browsing sessions and accurately estimate various web QoE metrics from network traces. Our classification approach has 20% higher precision than previous state-of-the-art. Further, we analyzed the impact of various radio network factors on web QoE. We observed that web QoE is particularly sensitive to IRAT handovers, ECNO and load in the cell. Moreover, we identified radio metrics that are often considered important, but show negligible impact on web QoE, such as average radio link data rate, soft handovers, and inter-frequency handovers. Finally, we developed *accurate* and *intuitive* machine learning models that can predict various web QoE metrics from *network factors alone*. Our models can be used by network operators to monitor web QoE using standard radio network metrics alone and prioritize improvement of network factors that have the highest impact.

Chapter 5

Conclusions and Future Work

In this dissertation we showed that applying large scale data analytics is a step forward towards solving some of the main challenges faced by the various players in the content delivery. We showed that large scale data analytics and machine learning algorithms can be used as an effective tool to characterize user behavior in the wild to inform various content delivery system design decisions. This chapter concludes the dissertation with a summary of the approach and contributions followed by a discussion of the lessons learned and remaining open problems in this space.

5.1 Summary of Approach

This dissertation presented the following thesis: *It is possible for different players to use big data analytics on user behavior data for learning predictive models that can be used to improve content delivery even when the data collected does not explicitly contain user behavior information.*

For substantiating the statement, this dissertation showed initial promise on using large scale data analytics of user behavior in the wild for addressing three main challenges faced by the players in the content delivery ecosystem today—(i) exponentially increasing traffic, (ii) increasing user expectations for quality of content and (iii) rise of mobile traffic. Below, we summarize these challenges and highlight our main contributions.

5.1.1 CDN Resource Management for Handling Increasing Traffic

Key Insight and Motivation: Traffic on the Internet has been on a rise mainly because of increasing video viewership. Market predictions suggest that video will comprise 90% of the traffic on the Internet by 2015. Because of the increasing traffic, there have been signs that the Content Delivery Network (CDN) infrastructure is being stressed by the ever-increasing amounts of video traffic. To meet these growing demands, the CDN infrastructure must be designed, provisioned and managed appropriately. In this context, federated telco-CDNs and hybrid P2P-CDNs are two content delivery infrastructure designs aimed at handling the increasing workload that have gained significant industry attention recently.

Core Contributions: Using large scale data analytics on user behavior data, we observed several user access patterns that have important implications to these two designs in our unique dataset consisting of 30 million video sessions spanning around two months of video viewership from two large Internet video providers. These include partial interest in content, regional interests, temporal shift in peak load and patterns in evolution of interest. We analyzed the impact of our findings on these two designs by performing a large scale measurement study. Surprisingly, we found significant amount of synchronous viewing behavior for Video On Demand (VOD) content, which makes hybrid P2P-CDN approach feasible for VOD and suggested new strategies for CDNs to reduce their infrastructure costs. We also found that federation can significantly reduce telco-CDN provisioning costs by as much as 95%. Using large scale data analytics and simulation studies we were able to inform several decisions in content provisioning and also develop models.

5.1.2 Predictive Model for Improving Video Quality of Experience

Key Insight and Motivation: Users expectation for quality of content has been on an increase. At the same time, improving users' quality of experience (QoE) is crucial for sustaining the advertisement-based and subscription-based revenue models that enable the growth of Internet video. Despite the rich literature on video and QoE measurement, our understanding of Internet video QoE is limited because of the shift from traditional methods of measuring video quality (e.g., Peak Signal-to-Noise Ratio) and user experience (e.g., opinion scores). These have been replaced by new quality metrics (e.g., rate of buffering, bitrate) and new engagement-centric measures of user experience (e.g., viewing time and number of visits).

Core Contributions: We identified two key requirements for a useful QoE model: (1) it has to be tied in to observable user engagement and (2) it should be actionable to guide practical system design decisions. Achieving this goal was challenging because the quality metrics are interdependent, they have complex and counter-intuitive relationships to engagement measures, and there are many external factors that confound the relationship between quality and engagement (e.g., type of video, user connectivity). To address these challenges, we presented a data-driven approach to model the metric interdependencies and their complex relationships to engagement, and proposed a systematic framework to identify and account for the confounding factors. Using a simulation study, we showed that a delivery infrastructure that uses our proposed model to choose CDN and bitrates can achieve more than 20% improvement in overall user engagement compared to strawman approaches.

5.1.3 Predictive Analytics for Extracting and Monitoring Cellular Web QoE

Key Insight and Motivation: Mobile traffic has been on an increase over the past few years. Recent studies have shown that web browsing is one of the most prominent cellular applications. However, mobile web browsing QoE is not as good as their wireline counterpart primarily because cellular networks are not as well designed for the web as wireline networks. It is therefore important for cellular network operators to understand how radio network characteristics (such

as signal strength, handovers, load, etc.) influence users' web browsing Quality-of-Experience (web QoE) in order to make infrastructure changes to improve web QoE. Also, understanding the relationship between web QoE and network characteristics is a pre-requisite for cellular network operators to detect when and where degraded network conditions actually impact web QoE. Unfortunately, cellular network operators do not have access to detailed server-side or client-side logs to directly measure web QoE metrics, such as abandonment rate and session length.

Core Contributions: We first devised a machine-learning-based mechanism to infer web QoE metrics from network traces accurately. We then presented a large-scale study characterizing the impact of network characteristics on web QoE using a month-long anonymized dataset collected from a major cellular network provider. Our results showed that improving signal-to-noise ratio, decreasing load and reducing handovers can improve user experience. We found that web QoE is very sensitive to inter-radio-access-technology (IRAT) handovers. We further found that higher radio data link rate does not necessarily lead to better web QoE. Since many network characteristics are interrelated, we also used machine learning to accurately model the influence of radio network characteristics on user experience metrics. This model can be used by cellular network operators to prioritize the improvement of network factors that most influence web QoE.

5.1.4 Summary of Thesis Contributions

Simple data analytics for informing system design decisions: We showed that even simple data analytics of user behavior can be useful for informing system design decisions towards improving content delivery. For instance, in the CDN resource management study, we observed several users who quit a session after watching the video for a few minutes (early quitters). We leveraged this observation for suggesting a better design of hybrid P2P-CDN architecture. Similarly, other observations made in that study, such as synchronous viewing behavior for VOD, temporal shift, regional interest in content have important implications to CDN resource management.

Machine learning for building predictive models: We also showed that large-scale data analytics, specifically machine learning tools can be used to develop predictive models of users QoE. For instance, we built a predictive model for Internet video QoE and performed simulation studies that showed initial promise that it be used for better bitrate and CDN selection. Similarly, we also built a predictive model for web browsing QoE that can help mobile network operators prioritize which cellular network factors need to be improved for better content delivery.

Extracting user behavior when it is not explicitly available: Not all players in the content delivery ecosystem have access to detailed client-side or server-side logs. For instance, network operators do not have access to such detailed logs to directly correlate various network factors to web QoE metrics. However, even in such a scenario, we showed that we can develop machine learning algorithms to extract these metrics from raw network traces.

5.2 Lessons Learned

Machine Learning is not a silver bullet: To many practitioners who throw machine learning (ML) toolkits and libraries at their data hoping to derive insights, machine learning often appears as a black box that magically gives answers to queries. While some off-the-shelf ML algorithms indeed work well for many common problems, it is more likely that they will not perform well—especially if they are applied without understanding the nature of the algorithm and the dataset. Even after picking the right algorithm, analyzing its performance is also critical for uncovering dependencies between parameters, which can then be accounted for to build a more robust model.

For real-world predictive tasks, picking the correct ML algorithm(s) is crucial and consequently non-trivial. The algorithm chosen should be expressive enough to capture all the relationships and dependencies between various parameters, and usually, there is no straightforward approach or cheat sheet for picking the best algorithm. However, performing preliminary analytics on the dataset can be very useful in *eliminating* some classes of ML algorithms.

For instance, in the predictive video QoE study, we saw that the quality metrics were interdependent on each other eliminating Naive Bayes as a possible algorithm for building a model. Similarly, we saw non-monotonic and non-linear relationships between quality metrics and engagement which is indicative that techniques like linear regression were not the best fit for modeling Internet video QoE. We finally chose decision trees since this algorithm does not have any of the above constraints. As another example, for the cellular network study, our correlation analysis showed that most of the network characteristics have a linear relationship with the QoE metrics. Therefore, we observed that linear regression techniques were able to build accurate models in this scenario.

Even after we identified the correct algorithm, we had to perform further analysis on the generated model to identify hidden dependencies. For example, the initial model for Internet video QoE was not highly accurate, but we noticed that there were external factors that affected user engagement. Identifying and accommodating for these factors led to a much more accurate model. Similarly, eliminating early quitters from the modeling led to further improvement in accuracy. In the cellular network study, noticing the differences between websites and the diurnal patterns of quality metrics and incorporating them into the model led to improvements.

To summarize, blindly learning models by trying various machine learning algorithms often produces a suboptimal model. Understanding the domain, the data and incorporating patterns in them can greatly help improve the performance of models that we build.

Obtaining ground truth and hidden confounding factors can be hard: One of the chief problems we faced in some sections of this thesis was the lack of good-quality ground truth on users quality of experience. We chose to use user engagement (e.g., play time of video, number of clicks during web session) as a measure for user quality of experience based on the assumption that if the user is not satisfied with the quality, she may terminate the session early. Further engagement metrics (e.g., number of clicks, play time etc) are objective and easier to capture/estimate at a large scale as opposed to the more subjective quality of experience metrics (e.g., opinion score, user rating); Also, engagement metrics can be directly translated into revenue for the players (e.g., higher play time, more web pages visited implies more ad impressions). However, engagement is not the best proxy for quality of experience. For instance, in the

video QoE study, we use fraction of video viewed as the engagement metric. However, a user could have left the session early not because of a quality issue —the session may have stopped due to a power outage, or perhaps even the user stopping the video to answer a phone call.

It is sometimes inherently hard, and sometimes even impossible, to obtain engagement metrics. In the Internet video QoE study, we were able to get good quality ground truth on engagement as we had access to client-side data collected by instrumentation libraries running on the clients. However, in the cellular network study, we had no way to tell whether a HTTP request was a user click or if it was an automated request for an embedded object. User clicks formed the basis of many of our web QoE metrics such as abandonment, number of user clicks during the session etc. In this case, we had to develop a click detection algorithm using URL text mining to predict what is and is not a user click.

Similarly, another issue that we repeatedly encountered in the thesis is the presence of external confounding factors that need to identified and incorporated into the model. We developed techniques to identify these confounding factors. However, there might be more hidden factors that were not captured in our dataset. One such example is user interest in content. For instance, a user would have quit the session early because she was not interested in the content and not because of a quality issue. Another such potential confounding factor is difference in users expectations for quality. While user A might prefer higher bitrate, user B might be more sensitive towards bit rate switching. Identifying these patterns from data and incorporating them into the model can also further improve prediction accuracies.

Generated models need updates for changing scenarios: While the broad techniques and the lessons learned from this research are general and can be applied on a similar dataset, the exact models learned in our research may not be applicable on a different content provider's data. The predictive models built by our chosen algorithms are not universal and only represent a classifier learned from the specific training set. This is because the models learn very specific parameters set by the specific content provider such as the bitrates they use, the buffer sizes they set, and so on. We believe that each content provider's data represents their unique video catalog and viewing patterns, so it is best to re-train the model using data and parameters specific to a single content provider.

Even within data generated by the same content provider, when the nature of data to be classified changes—such as video viewing patterns of users may change during a holiday season—the model will also need to be updated. Other parameters such as video bitrates may change or increase over time even for a given content provider. Similarly, for the cellular study, the models will need to be re-learned with the adoption of new technologies such as LTE and 5G.

For maintaining up-to-date models for data from a single content provider, there are several potential approaches. and identifying the best approach would depend on the specific problem domain. One option is to train the model with data not just for 3 months as we did in the QoE study, but using data for a whole year. However, this also increases the the time and space complexity of the ML training phase. Several ML algorithms have super linear time/space complexity and training these models will be difficult. An active area of research today is in developing parallellizable ML algorithms [17]: these techniques would permit training ML models using very large data sets using the combined resources of a compute cluster.

A second method for updating models is to train several different models by partitioning the

data upon depending on some characteristic, e.g., the month of year. However, this approach risks losing any interdependencies that exist between data points from different partitions. A third approach is to retrain the model periodically using, for example, the last X months of data. Although updating models may not be an expensive operation, picking the right method and understanding how frequently these models need an update in an appealing area for future work.

5.3 Future Work

The research in this dissertation suggests various avenues for future work in large-scale data analytics for content delivery. We present a few research directions below that may be good avenues for future work.

5.3.1 Improved techniques to re-learn and refresh models

As discussed earlier, models may need to be re-learned and updated when the data features change to maintain a high accuracy of classification, especially for live deployment scenarios. This requirement poses several challenges in the machine learning domain

The sometimes exponential complexity of training and classification make it prohibitively expensive to train on larger datasets. Most typical machine learning algorithms assume a shared memory address space and are hence not inherently parallelizable. In our video work, we trained on 3 months of data, and in the cellular study, we trained with data for 1 month; these timeframes were partly influenced by the physical memory limit of the system used to run the training code.

Network data is growing tremendously each year, and with ubiquity of Internet-connected devices, it may not be long before a system can handle training only a single day's worth of data. Sampling data might allow a longer timeframe to be trained, but this may miss important data points. The solution—parallelized machine learning algorithms [17]—is an active area of research, and further research in this area with a focus on network data will prove very valuable.

Creating ensemble models is another avenue that can be explored to improve models. Ensemble learning uses multiple learning algorithms to obtain a better resultant performance than that of any algorithm applied alone. Because there are several algorithms that may work on the same dataset with varying accuracies, applying an ensemble learning technique such as bagging, boosting, or random forests [101] may generate a better overall classifier.

Finally, predictive performance can be improved by identifying more confounding factors, personalizing models, and incorporating these during training.

5.3.2 Fine-grained video quality metrics using intra-session analysis

A promising avenue for research is to further investigate the dynamics of quality metrics within a single session rather than across sessions. The quality metrics that we looked at in our study were more "coarse-grained" session-level aggregate metrics such as average bitrate, buffering ratio, etc. Identifying and incorporating more fine-grained metrics such as bitrate at the beginning of the session ("initial bitrate"), the time of the buffering event, etc. may lead to better models. Intuitively, a fine-grained metric such as the time of buffering event may be quite significant to a

user's engagement: if the video buffers several times at the beginning, the user may be frustrated and quit early, whereas they may be more tolerant of buffering a bit further into the video. Such insights may also inform system design: users may prefer a video to begin playing with minimal buffering even if the bitrate is lower, and system designers can optimize their infrastructure to account for these tradeoffs.

5.3.3 Web QoE model for Cellular Network Operations

The Web QoE model predicts user engagement directly from network parameters such as signal strength, handovers, throughput, and so on. The model may also be used to build a real-time alerting tool to detect widespread low QoE in a region as predicted by our model. Previously, network operators would have had to perform deep packet inspection on HTTP traffic to understand engagement, but our model makes it easier to predict various engagement metrics directly from easily-gathered network factors.

The alerting tool may need further improvements beyond our QoE model, for example, it may need to actively monitor and benchmark the engagement of users on the carrier network across time and geographical region. This anomaly detection application may also benefit from even minor improvements in accuracy for the QoE model since it would directly translate to fewer false positive alerts and more correct alerts.

5.3.4 Predictive Analytics for Other Aspects of Content Delivery

Large-scale data analytics can be used to improve several other aspects of content delivery. The focus of this thesis was to perform data analytics of user behavior to improve content delivery. However, data analytics of network throughput can be useful to predict which parts of the network pose bottlenecks. This can be useful for network operators of improving resource provisioning. Building a real time monitoring system using such a model can be useful for traffic management and rerouting traffic to avoid congested paths.

Other interesting problems include picking the best content to be cached at local servers, real time prediction of throughput within a session etc.

Bibliography

[1] 100Gbps Ethernet Task Force. http://www.ieee802.org/3/ba/.

[2] Use RTMFP for developing real-time collaboration applications. http://labs.adobe.com/technologies/cirrus/.

[3] Akamai investor summit 2013. http://www.akamai.com/dl/investors/2013_ir_summit_presentation.pdf,.

[4] Akamai NetSession. http://www.akamai.com/client/,.

[5] Driving Engagement for Online Video. http://goo.gl/pO5Cj,.

[6] Census Bureau Divisioning. http://www.census.gov/geo/www/us_regdiv.pdf.

[7] Buyer's Guide: Content Delivery Networks. http://goo.gl/B6gMK.

[8] Cisco Report on CDN Federation - Solutions for SPs and Content Providers To Scale a Great Customer Experience. .

[9] Cisco visual networking index: Global mobile data forecast update 2013-2018. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html,.

[10] Cisco study. http://goo.gl/tMRwM, .

[11] Conviva. http://www.conviva.com.

[12] MPEG - DASH. http://dashif.org/mpeg-dash/.

[13] DCCP. https://tools.ietf.org/html/rfc4340.

[14] Firebug. getfirebug.com.

[15] Amazon down for 30 minutes. http://www.forbes.com/sites/kellyclay/2013/08/19/amazon-com-goes-down-loses-66240-per-minute/,.

[16] Google page speed. https://developers.google.com/speed/pagespeed/,.

[17] Graphlab. http://graphlab.com/.

[18] Hadoop. http://hadoop.apache.org/.

[19] Hdfs. http://hadoop.apache.org/docs/r0.18.0/hdfs_design.pdf.

[20] Hive. http://hive.apache.org/.

[21] For Impatient Web Users, an Eye Blink Is Just Too Long to Wait. http://cisephysics.homestead.com/files/BallsPitchedToHomePlate.pdf.

[22] Mapreduce. http://research.google.com/archive/mapreduce.html.

[23] P.800 : Methods for subjective determination of transmission quality. http://www.itu.int/rec/T-REC-P.800-199608-I/en,.

[24] P.910 : Subjective video quality assessment methods for multimedia applications. http://goo.gl/QjFhZ,.

[25] Mail service costs Netflix 20 times more than streaming. http://www.techspot.com/news/42036-mail-service-costs-netflix-20-times-more-than-streaming.html.

[26] Next-Generation Optical Transport Networks. http://www.jdsu.com/ProductLiterature/next-generation-optical-transport-network-white-pape pdf.

[27] NFL What are HQ videos? http://www.nfl.com/help/faq.

[28] Ooyala. http://www.ooyala.com/.

[29] Peak signal to noise ratio. PSNRWikipedia.

[30] Wcdma - network planning and optimization. http://www.qualcomm.com/media/documents/files/wcdma-network-planning-and-optimization.pdf,.

[31] Quantcast website ranking. https://www.quantcast.com/top-sites/US,.

[32] RTSP. http://www.ietf.org/rfc/rfc2326.txt.

[33] Scikit-learn. http://scikit-learn.org/stable/.

[34] SCTP. https://www.ietf.org/rfc/rfc2960.txt.

[35] Microsoft Smooth Streaming. http://goo.gl/6JOXh.

[36] Spark. https://spark.apache.org/.

[37] SPEC philosophy. http://www.spec.org/spec/#philosophy.

[38] Text classification and naive bayes. http://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html.

[39] Turbobytes: How it works. http://www.turbobytes.com/products/optimizer/.

[40] Vqeg. http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx.

[41] WebRTC 1.0: Real-time Communication Between Browsers. http://www.w3.org/TR/webrtc/.

[42] Weka. http://www.cs.waikato.ac.nz/ml/weka/,.

[43] Data mining with weka. http://goo.gl/YD9Awt,.

[44] Using Predictive Prefetching to Improve World Wide Web Latency. In *SIGCOMM Computer Communivation Review*, 1996.

[45] Experiences in a 3G Network: Interplay between the Wireless Channel and Applications. In *MOBICOM*, 2008.

[46] Tackling bufferbloat in 3G/4G networks. In *IMC*, 2012.

[47] An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance. In *SIGCOMM*, 2013.

[48] B. Niven-Jenkins, F. L. Faucheur, and N. Bitar. Content distribution network interconnection (CDNI) problem statement.

[49] K. Chen, C. Huang, P. Huang, C. Lei. Quantifying Skype User Satisfaction. In *Proc. SIGCOMM*, 2006.

[50] L. Plissonneau and E. Biersack. A Longitudinal View of HTTP Video Streaming Performance. In *Proc. MMSys*, 2012.

[51] Henrik Abrahamsson and Mattias Nordmark. Program popularity and viewer behavior in a large TV-on-Demand system. In *IMC*, 2012.

[52] S Akhshabi, A Begen, and C Dovrolis. An Experimental Evaluation of Rate Adaptation Algorithms in Adaptive Streaming over HTTP. In *MMSys*, 2011.

[53] Saamer Akhshabi, Lakshmi Anantakrishnan, Constantine Dovrolis, and Ali C. Begen. What Happens when HTTP Adaptive Streaming Players Compete for Bandwidth? In *Proc. NOSSDAV*, 2012.

[54] Robert H Allen and Ram D Sriram. The Role of Standards in Innovation. *Elsevier: Technology Forecasting and Social Change*, 2000.

[55] David Applegate, Aaron Archer, Vijay Gopalakrishnan, Seungjoon Lee, and Kadangode K. Ramakrishnan. Optimal Content Placement for a Large-Scale VoD System. In *Proc. CoNext*, 2010.

[56] B. Cheng, L. Stein, H. Jin, and Z. Zheng. Towards Cinematic Internet Video-On-Demand. In *Proc. Eurosys*, 2008.

[57] Athula Balachandran, Vyas Sekar, Aditya Akella, Srinivasan Seshan, Ion Stoica, and Hui Zhang. A quest for an internet video quality-of-experience metric. In *Hotnets*, 2012.

[58] Athula Balachandran, Vyas Sekar, Aditya Akella, Srinivasan Seshan, Ion Stoica, and Hui Zhang. Developing a Predictive Model of Quality of Experience for Internet Video. In *SIGCOMM*, 2013.

[59] Hari Balakrishnan, Venkata N. Padmanabhan, Srinivasan Seshan, Mark Stemm, and Randy H. Katz. TCP Behavior of a Busy Internet Server: Analysis and Improvements. In *INFOCOM*, 1998.

[60] Paul Barford and Mark Crovella. Generating Representative Web Workloads for Network and Server Performance Evaluation. In *SIGMETRICS*, 1998.

[61] Anna Bouch, Allan Kuchinsky, and Nina Bhatti. Quality is in the Eye of the Beholder: Meeting Users Requirements for Internet Quality of Service. In *CHI*, 2000.

[62] Michael Butkiewicz, Harsha V. Madhyastha, and Vyas Sekar. Understanding Website Complexity: Measurements, Metrics and Implications. In *IMC*, 2011.

[63] C. Huang, A. Wang, J. Li, and K. W. Ross. Understanding Hybrid CDN-P2P: Why Limelight Needs its Own Red Swoosh. In *Proc. NOSSDAV*, 2008.

[64] C. Huang, J. Li, and K. W. Ross. Can Internet Video-on-Demand be Profitable? In *Proc. SIGCOMM*, 2007.

[65] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proc. IMC*, 2007.

[66] S Chikkerur, V Sundaram, M Reisslein, and L J Karam. Objective video quality assessment methods a classification review and performance comparison. In *IEEE Transactions on Broadcasting*, 2011.

[67] Brian Clifton. *Advanced Web Metrics with Google Analytics*. John Wiley and Sons, 2012.

[68] Nicola Cranley, Philip Perry, and Liam Murphy. User perception of adapting video quality. *International Journal of Human-Computer Studies*, 2006.

[69] Heng Cui and Ernst Biersack. On the Relationship between QoE and QoS for Web Sessions. In *Eurecom Research Report*, 2012.

[70] D. Rayburn. Telcos and carriers forming new federated cdn group called ocx (operator carrier exchange). June 2011. StreamingMediaBlog.com.

[71] Houtao Deng, George Runger, and Eugene Tuv. Bias of importance measures for multivalued attributes and solutions. In *ICANN*, 2011.

[72] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Antony Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. Understanding the impact of video quality on user engagement. In *Proc. SIGCOMM*, 2011.

[73] Jeffrey Erman, Alexandre Gerber, K.K. Ramakrishnan, Subhabrate Sen, and Oliver Spatscheck. Over the top video: The gorilla in cellular networks. In *IMC*, 2011.

[74] Jairo Esteban, Steven Benno, Andre Beck, Yang Guo, Volker Hilt, and Ivica Rimac. Interactions Between HTTP Adaptive Streaming and TCP. In *Proc. NOSSDAV*, 2012.

[75] Bin Fan, David Andersen, Michael Kaminsky, and Konstantina Papagiannaki. Balancing Throughput, Robustness, and In-Order Delivery in P2P VoD. In *Proc. ACM CoNEXT*, 2010.

[76] Alessandro Finamore, Marco Mellia, Maurizio Munafo, Ruben Torres, and Sanjay G. Rao. Youtube everywhere: Impact of device and infrastructure synergies on user experience. In *Proc. IMC*, 2011.

[77] Dennis F. Galletta, Raymond Henry, Scott McCoy, and Peter Polak. Web Site Delays: How Tolerant are Users? In *Journal of the Association of Information Systems*, 2004.

[78] Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. Large-Scale Sentiment Analysis for News and Blogs. In *ICWSM*, 2007.

[79] Alon Halevy, Peter Norvig, and Fernado Pereira. The unreasonable effectiveness of data.

In *IEEE Intelligent Systems*, 2009.

[80] Ahmed Hassan, Yang Song, and Li wei He. A Task Level Metric for Measuring Web Search Satisfaction and its Application on Improving Relevance Estimation. In *CIKM*, 2011.

[81] Xiaojun Hei, Chao Liang, Jian Liang, Yong Liu, and Keith W. Ross.

[82] L Huang, J Jia, B Yu, B G Chun, P Maniatis, and M Naik. Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression. In *Proc. NIPS*, 2010.

[83] Yan Huang, Dah-Ming Chiu Tom Z. J. Fu, John C. S. Lui, and Cheng Huang. Challenges, Design and Analysis of a Large-scale P2P-VoD System. In *Proc. SIGCOMM*, 2008.

[84] Sunghwan Ihm and Vivek S Pai. Towards Understanding Modern Web Traffic. In *IMC*, 2011.

[85] Oliver Spatscheck Walter Willinger John Otto, Fabian Bustamante. A Balancing Act: Content Distribution, Networks and Their Users. In *SIGMETRICS*, 2014.

[86] A Khan, Lingfen Sun, and E Ifeachor. Qoe prediction model and its applications in video quality adaptation over umts networks. In *IEEE Transactions on Multimedia*, 2012.

[87] S. Shunmuga Krishnan and Ramesh K. Sitaraman. Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs. In *IMC*, 2012.

[88] S. Shunmuga Krishnan and Ramesh K. Sitaraman. Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs. In *IMC*, 2012.

[89] Shunmuga Krishnan and Ramesh Sitaraman. Video Stream Quality Impacts Viewer Behavior: Infering Causality Using Quasi-Experimental Designs. In *IMC*, 2012.

[90] Bo Li, Susu Xie, Yang Qu, and Keung G.Y. Inside the New Coolstreaming: Principles, Measurement and Performance Implications. In *INFOCOMM*, 2008.

[91] Zhenyu Li, Jiali Lin, Marc-Ismael Akodjenou-Jeannin, Gaogang Xie, Mohamed Ali Kaafar, Yun Jin, and Gang Peng. Watching video from everywhere: a study of the pptv mobile vod system. In *IMC*, 2012.

[92] Bing Liu, Minqing Hu, and Wynne Hsu. Intuitive Representation of Decision Trees Using General Rules and Exceptions. In *Proc. AAAI*, 2000.

[93] Chao Liu, Ryen W. White, and Susan Dumais. Understanding Web Browsing Behaviors through Weibull Analysis of Dwell Time. In *SIGIR*, 2010.

[94] Harry Liu, Ye Wang, Yang Richard Yang, Alexander Tian, and Hao Wang. Optimizing Cost and Performance for Content Multihoming. In *to appear in SIGCOMM 2012*, 2012.

[95] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang. A Case for a Coordinated Internet Video Control Plane. In *Proc. SIGCOMM*, 2012.

[96] Xi Liu, Florin Dobrian, Henry Milner, Junchen Jiang, Vyas Sekar, Ion Stoica, and Hui Zhang. A case for a coordinated internet video control plane. In *SIGCOMM*, 2012.

[97] Bruce A Mah. An Empirical Model of HTTP Network Trafc. In *INFOCOM*, 1997.

[98] Ajay Mahimkar, Zihui Ge, Aman Shaikh, Jia Wang, Jennifer Yates, Yin Zhang, and

Qi Zhao. Towards Automated Performance Diagnosis in a Large IPTV Network. In *Proc. SIGCOMM*, 2009.

[99] V Menkvoski, A Oredope, A Liotta, and A C Sanchez. Optimized online learning for qoe prediction. In *BNAIC*, 2009.

[100] Leo A. Meyerovich and Rastislav Bodk. Fast and Parallel Webpage Layout. In *WWW*, 2010.

[101] Tom Mitchell. *Machine Learning*. McGraw-Hill.

[102] Jeffrey C. Mogul. The Case for Persistent-Connection HTTP. In *SIGCOMM*, 1995.

[103] Ricky K. P. Mok, Edmond W. W. Chan, Xiapu Luo, and Rocky K. C. Chang. Inferring the QoE of HTTP Video Streaming from User-Viewing Activities . In *SIGCOMM W-MUST*, 2011.

[104] Tongqing Qiu, Zihui Ge, Seungjoon Lee, Jia Wang, Qi Zhao, and Jun Xu. Modeling channel popularity dynamics in a large IPTV system. In *Proc. SIGMETRICS*, 2009.

[105] R. Powell. The federated cdn cometh. May 2011. TelecomRamblings.com.

[106] S. Guha, S. Annapureddy, C. Gkantsidis, D. Gunawardena, and P. Rodriguez. Is High-Quality VoD Feasible using P2P Swarming? In *Proc. WWW*, 2007.

[107] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A Bayesian Approach to Filtering Junk E-Mail. In *AAAI*, 1998.

[108] M. Zubair Shafiq, Lusheng Ji, Alex X. Liu, Jeffrey Pang, and Jia Wang. Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network . In *INFOCOM*, 2012.

[109] F. Donelson Smith, Felix Hernndez Campos, Kevin Jeffay, and David Ott. What TCP/IP Protocol Headers Can Tell Us About the Web. In *SIGMETRICS*, 2001.

[110] Han Hee Song, Zihui Ge, Ajay Mahimkar, Jia Wang, Jennifer Yates, Yin Zhang, Andrea Basso, and Min Chen. Qscore Proactive Service Quality Assessment in a Large IPTV System. In *Proc. IMC*, 2011.

[111] Srikanth Sundaresan, Nick Feamster, Renata Teixeira, and Nazanin Magharei. Measuring and Mitigating Web Performance Bottlenecks in Broadband Access Networks. In *IMC*, 2013.

[112] Thomas Tullis and William Albert. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008. ISBN 0123735580, 9780123735584.

[113] Mark Watson. Http adaptive streaming in practice. In *MMSys - Keynote*, 2011.

[114] I H Witten and E Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2000.

[115] C Wu, B Li, and S Zhao. Diagnosing Network-wide P2P Live Streaming Inefficiencies. In *Proc. INFOCOM*, 2009.

[116] Wei Xu, Ling Huang, Armando Fox, David Patterson, and Michael Jordan. Detecting large-scale system problems by mining console logs. In *Proc. SOSP*, 2009.

[117] H Yin et al. Inside the Bird's Nest: Measurements of Large-Scale Live VoD from the 2008 Olympics.

[118] Hao Yin, Xuening Liu, Tongyu Zhan, Vyas Sekar, Feng Qiu, Chuang Lin, Hui Zhang, and Bo Li. Design and Deployment of a Hybrid CDN-P2P System for Live Video Streaming: Experiences with LiveSky. In *Proc. ACM Multimedia*, 2008.

[119] H Yu, D Zheng B Y Zhao, and W Zheng. Understanding User Behavior in Large-Scale Video-on-Demand Systems. In *Proc. Eurosys*, 2006.

[120] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of YouTube recommendation system on video views. In *Proc. IMC*, 2010.